International Journal of Computer Sciences and Engineering
Vol.**11**, Issue **4**, pp.**07-13**, April **2023**
ISSN: 2347-2693 (Online)
Available online at: www.ijcseonline.org

ISSN: 2347-2693 (E)

Research Paper

Sign and Voice Translation using Machine Learning and Computer Vision

Nandini [1*], Avni Verma[2], Sandeep Kumar[3]

[1,2,3]Department of Computer Science and Technology, Sharda School of Engineering & Technology, Sharda University, Greater

**Corresponding Author: nandini.sharmaa007@gmail.com*

Received: 27/Feb/2023; **Accepted**: 08/Apr/2023; **Published**: 30/Apr/2023. | **DOI**: https://doi.org/10.26438/ijcse/v11i4.713

Abstract: Sign and voice translation is a critical tool for individuals who cannot hear or speak , or for those who speak different languages. Machine learning techniques have been increasingly used to find or improve the accuracy and efficiency of sign and voice translation systems. These systems make use of machine learning models to analyze and interpret sign language or speech and translate them into written or spoken language. Machine learning models can recognize patterns in sign language gestures or speech, and convert them into text or speech output. The model's accuracy is dependent on the quality of its training data and the complexity of the model architecture. Recent improvisation in machine learning has increased the performance of sign and voice translation systems, enabling them to recognize more complex gestures and accents. Overall, the use of machine learning in sign and voice translation has the potential to improve the accessibility of information and communication for individuals who are deaf or hard of hearing, or for those who speak different languages. However, there is still much room for improvement, and ongoing research and development are needed to optimize the performance of these systems

Keywords: Computer Vision, Recognition of Sign Language, Hand Gesture Recognition, Features Extraction

1. Introduction

Nowadays, it is quite challenging for people who cannot speak or hear to interact with normal people who can talk because they interact by sign language, which has not evolved from spoken languages; their vocabulary and grammatical structures seem to be unique. People like most of us won't be capable of communicating with normal folks if and until they learn sign language. A total of 360 million beings all throughout the world, comprising a total of 328 million adults and 32 million of youngsters, suffer from hearing impairment, as per the World Federation of the Deaf(WFD) assessment. There are over 300 sign languages being used worldwide[3]. Additionally, single and double hand articulation is used to convey significant signs. A word in sign language can be expressed only with a single gesture, which would be known as compact expression. Since sign language alphabets vary throughout sign languages, recognition of sign is a challenging problem.

For instance The alphabets used in American Sign Language (ASL) distinguished greatly from those used in Indian Sign Language or Italian Sign Language.. Through the utilization of visual sign patterns, emotions could be expressed in sign language. The deaf community uses sign language commonly, yet it is not widely accepted. There is a language gap when people who are deaf try to convey their thoughts, opinions and hearing with the general public.

Currently two organizations focus heavily on human interpreters, which makes them costly and time-consuming [11]. The underlying structure of the signs used by the deaf is similar to those of spoken words. Likewise to how tens of Millions of English words are composed by just a couple of different sounds, yet SL signals can be created by a restricted amount of gesturing features. As a result, signs are not whole movements but rather may be analyzed as a collection of linguistically important elements. SLs are made up of the following indivisible characteristics, much as spoken languages:

- It is crucial to recognize manual features such as hand shape, postures, movement, and palm and finger orientation while even the simplest of human gestures may impose several grammatical and semantic readings depending on many factors [6].
- The speed at which a certain texture is applied might offer up the dumb's attitude or indicate a specific interpretation. Rather than using two interpretations to indicate "move fast," deaf and dumb might permit the relevant signs more quickly [6]
- Since there haven't been many formal standardization efforts made for most SLs up to this point, signers from the same nation can be distinguished while a particular gloss is being executed [4].
- Even as there haven't been many formalized standards attempts for the majority of sign language up to this date,

deaf and dumbs from the same country can be identified while a specific interpretation is carried out.

In this research, a suggested intelligent system for human sign language based speech synthesis from face and hand gestures. This kind of two way communication system uses a camera to receive sign language and a microphone for voice recognition in real time from deaf and dumb. Each photo frame is divided into the two-hands and head region which are then combined to form an image. Another way can be by transforming the voice into signs by using the dataset of signs. This problem is representing a significant challenge from the point of view of computer vision because of a number of characteristics, such as [4]:

- Environment-related factor(such as backdrop, lighting effect, camera location)
- Occlusion(when a hand, with all its fingers may obstruct the part of view)
- Recognising signs boundaries(where one sign finishes and another one starts)
- Coarticulation (where sign is impacted by previous or following sign)

Signed language is a form of sign language used by the deaf and dumb as their native languages. Rather than using acoustically transmitted Sign speech, rather than mechanically delivered patterns of sounds, is a way to dynamically symbolize somebody's emotions. People find it difficult to speak with these people with special needs since learning Sign Language takes a lot of time, which results in a communication gap. As a result, we suggest an application that accepts live voice or audio recordings as input, translates them into text, and then displays the appropriate pictures or GIFs in Indian Sign Language. The person with disabilities would place their fingers in front of a webcam, and the camera would capture the motion of the hand before doing Principle Component Analysis (PCA) image processing. The exact image from the database will be recognised when the acquired coordinates are mapped with the previously recorded image. By keeping on this way, a person who is physically disabled will be capable of communicating the complete sentence. Afterwards, the sentence will be made into speech for everyone to hear. This approach would benefit the mutes since it would permit them to freely speak with one another. An oral obstruction is an illness. Restricts a person's ability to interact socially through speech and hearing. Numerous different forms of communication, such as sign language are used by those who can't speak or hear. Regardless of the fact that sign language is now frequently used, it's still difficult for people who don't know signs to communicate with the sign language users. Recent developments in technology have produced promising results in the movement and gesticulation detection domains using deep learning. Using this proposed approach, both temporal and spatial properties are extracted from the webcam inputs. This initiative makes an effort to close this gap by adding a low computational power computer to the communication medium, allowing sign language to be automatically gathered, then recognised and translated into speech for benefits of blind. Speech must be processed and converted to either sign language or text for screen display for benefit of those who are hard of hearing. An important research topic for such a system is the computerized recognition of sign language using image processing. This paper details an image processing method we developed specially for recognising the signs and speech in Indian sign language. The recognition of gestures made by hand recently has been the center of vigorous study. We want to build a robust, non-person reliant system that can recognise continuous sign language messages. We adopted a vision based method that doesn't need the use of specialized data collection equipment, such as hand gloves or motion detection devices. For a prototype, a simple webcam can work. A SLR system that has sensors based on earlier studies proved extremely uncomfortable and constrained for deaf and dumb [2, 5], Specialized equipment such as sensors were employed, which was also a costly choice. On the other hand , techniques based on computer vision use human unequipped hands without sensors or coloured gloves. As just one camera is used, computer vision based methods are more affordable and portable than sensor based ones. The most frequently used approach for hand tracking in computer vision based systems is background and skin color reduction. SLR systems based on computer vision usually deal with features extraction, including estimation of hand forms, boundary modeling, contour, gestures segmentation and gestures recognition.

2. Literature Survey

In the domain of machine learning, hand motion recognition is a challenging topic to address. The two fundamental criteria of classification strategies tend to be supervised and unsupervised. The SLR systems, that are based on these methods, can recognise moving and stationary hand signs. In 1991, Murakami and Taguchi released a study in which they used neural networks to recognise sign language for the first time. Several researchers developed cutting edge methods to assist the community of physically challenged people as a result of advancements in computer vision. Wang and Popovic created a real-time hand tracking technology utilizing pigmented gloves [18]. The K- Nearest Neighbors (KNN) approche was utilized to identify the color patterns of the hand gloves, although the system needs hand stream feeding. In the research results of Rekha et al.[13], Support Vector Mechanism (SVM) beat this approach. Kurdyumov et al.[9], Tharwat et al.[9] and Baranwal and Nandi[1] proposed that there are two distinct types of sign recognition: Continuous sentence recognition and isolated sign recognition. Similar modeling approaches exist in the SLR system for whole and subunit signs. Two methods that eventually contribute to subunit level of sign recognition and they are visual-descriptive and linguistic-focused. To construct a framework for alphabet subunit recognition, Elakkiya et al.[19] integrated SVM learning and boosting techniques. The classifier succeeded in recognizing 97.6% of alphabets, however it misjudges 26 of them. Ahmed and Aly[10] used PCA and local binary patterns to extract attributes from 23 distinct Arabic sign languages. While

receiving the accuracy of 99.97% of the sign language model the system was unable to identify the persistent grey-scale patterns of signs in the format of image because of the use of threshold operator. The challenge of hand gesture recognition in machine learning is comparatively difficult to address. In the early efforts to recognise the hand motions from the fragments of the photo, a traditional convolutional neural network was commonly used. R.Sharma et al.[15] used 80,000 different signs along with more than 500 images each to build a machine learning model. A training dataset of pre-processed pictures and videos are the part of their system technique for both hand-detection system and a motion recognition system. Before training a machine learning model, feature extraction was used to normalize the input data from the images. The images are flattened into fewer one -dimensional elements getting converted into gray level for enhanced object contour while maintaining a standardized resolution. The feature extraction approach aids in the extraction of specific pixel-level characteristics from pictures, which are then fed to CNN for faster learning and more accurate prediction. W. Liu et al have completed hand tracing in 2D and 3D space. A classification accuracy of nearly 98% was acquired using skin saliency, in which tones within a certain range were removed from improved feature extraction. All of the methods mentioned above show that models need a large dataset, a challenging methodology and extensive mathematical processing in order to recognise hand movement correctly with high accuracy. The gesture recognizing procedure relies heavily on image pre-processing. As a consequence, we decided to use Mediapipe, a google open-source framework which is proficient in accurately identifying human body parts, for our research.

The difficulty of ASL recognition in computer vision is not new. During the last 20 years, researchers have used a variety of classifiers, which we may generally categorize into classifiers based on linearity, artificial neural networks and Bayesian networks [17]. Although linear classifiers are convenient to use, they have to provide complicated feature extractions and preprocessing techniques to be efficient [2, 3, 4]. By implementing Karhunen-Loeve Transforms, Singha and Das were able to attain 96% accuracy on 10 classes of photographs showing one-handed actions. The dimensions are translated and rotated to construct a new coordinate structure based on the variance of the data points. The images are changed once they have been edge detected, manually cropped and subjected to a skin filter. To differentiate between the hand gestures like the thumbs-up, the index finger pointing left or right and numerals they use linear classifiers (no ASL). Sharma et al. [14] make use of a piecewise classifier such as Support Vector Machine and KNearest Neighbors to characterize every color frame after surround and noise elimination. Their creativity is the outcome of the contour trace they applied, accurate representation of hand outline. They obtained 62.3% accuracy with the help of SVM on the segregated color channel model. Non-verbal gestures in sign language are used by people to pass on their thoughts and feelings. Moreover, normal people have an extremely difficult time understanding sign language, so trained sign language translators are required to convey the words and thoughts of deaf and dumb. In order to select spatial attributes from the image for sign recognition, instructional and training period use a CNN model called inception [7]. After that, To select the temporal attributes from the images of signs by deploying the LSTM(Long Short- Term Memory) and RNN(Recurrent Neural Network) model via two techniques: using the softmax and pool layers of CNN's output, respectively[7]. A programme that can automatically identify the static alphabets hand signs in ASL [16].AdaBoost and Haar-like classifiers are two parallel connected approaches that are used to enhance it. A significant dataset was compiled for the training process in order to increase the accuracy of the model, and it delivered excellent results.This takes the live pictures as input and output as T. A variety of fresh potentials for Human Computer Interaction(HCI) have already been made possible by recent development depth sensors, specially the Kinect sensors [16]. Even though the Kinect sensor has made substantial improvements in human body tracing, face and motion detection and reliable hand gesture recognition it remains a work in progress. The hands are seen as a tiny identity with more complicated expressions than the whole body and are more sensitive to segmentation errors. As a result, it is particularly hard to recognize hand motions. By using the Kinect sensor to focus on a vigorous part-based hand motion detection system. Finger Earth Mover's Distance (FEMD), a revolutionary distance metric that could manage the irregular forms from the Kinect sensor, was suggested to evaluate variance between hand appearance which only matches the finger portions, not the entire hand, consequently making it simpler to differentiate between small variations in hand movements. A related piece of study uses a recurrent three dimensional convolutional neural network to perform dynamic hand gestures classifications and motion detection from multi-model data [20]. A technique for jointly segmenting vigorous hand motions from continuous depth, color and stereo-IR data stream. This system utilizes a recurrent 3-D CNN with connectionist temporal classification (CTC) to expand on the latest favorable result of CNN classifier for gesture recognition [12]. In a different study , machine learning methods and skin segmentation were used to construct an ASL translation. There, a color based automatic human skin segmentation method was applied. As a bivariate normal distribution in the CbCr plane, the distribution of skin color. In the process of extracting characteristics from the image , Convolutional Neural Networks (CNN) are used. Deep Learning technique is then utilized to train classifiers to recognise Sign Language.

3. Methodology

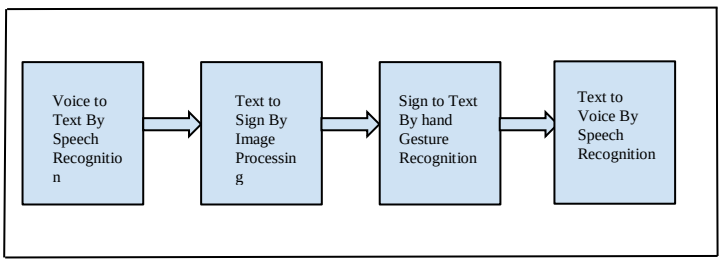

Fig. 1 : Complete Flow of working

3.1Convergence of Voice to Sign

3.1.1. Importing packages

The purpose of converting the spoken words into text, presenting ideas or receiving responses, a variety of libraries are used in this module including speech recognition. Perhaps few machines can be set up to react to these utter words.. Python interpreter acquires the ability to process images because of the Python Imaging Library (PIL). This library offers the capabilities for a variety of file formats, a useful internal representation and somewhat powerful image processing. A multidimensional array object with impressive results is available from Numpy, along with tools for working with the arrays. The foundational python module for scientific computing, to make it simple. The software is freely available [18]. Programming a GUI seems to be very easy and uncomplicated due to Python's EasyGUI module. EasyGUI is different from other presented GUI frameworks because it's not event-driven. Instead, all GUI interactions are initiated by simple function calls. Matplotlib is an optional choice for visualizing frames from films or photographs. Tkinter is the name of Python's built-in GUI library. The implementation of GUI applications is rapid and easy due to Python and Tkinter conjunction. Tkinter offers a useful object-oriented interface for the Tk GUI toolkit. Tkinter module import.

3.1.2. Converting The Voice to Text

Computational semantics is used to translate voice into words or text using voice recognition and Natural language Processing (NLP). Microphone will be used to acquire the voice inputs, which will then be analyzed by speech recognition software before converting to text. The text will be used for additional processing, such as translating that text into sign language.

3.1.3. Converting The Text to images

The dataset will be converted into GIFs or photos, then store the images according to their word name. The speech will now be converted into text, which will then be checked into the dataset. Does a term that is comparable to the transformed text? If so, the text-saving picture will be presented;otherwise, the show won't recognise it [10].

➢ If "bye" is detected as text inside predefined dictionary words, go on to the next option. Display the relevant GIFs for the phase.

➢ Instead, count the letters in the words or phrases. With a small delay in the activities, display the sentence visually.

➢ Up until the speech's completion, strictly adhere to step3's directions.

Fig. 2 : Convergence to text into Sign

3.2 Convergence of Sign to Voice

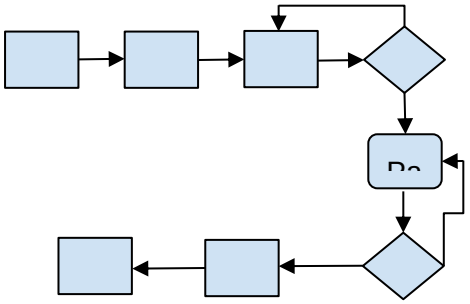

Fig. 3: Workflow for Converting Sign to voice

3.2.1. Image Preprocessing to acquire multi-hand landmarks

The development of cross-platform, multidimensional(video, audio, any time series data) applied ML pipeline is made possible by the MediaPipeline framework. The numerous human body recognition and tracing models provided by Midea Pipe were trained by Google's enormous and wide-ranging dataset for tracing. They behave as the structure of nodes, edges or markers, keeping monitor of significant locations on several body parts. Every coordinate point is normalized in three dimensions. Information flow may be easily modified and adapted due to a model developed by google developers using Tensorflow lite. Nodes on a graph, which makeup MediaPipe pipeline, are most commonly explained in pb text files. Most such elemental nodes are linked to c++ files. These files are an expansion of the mediapipe basic calculator class. This class ensures that it is linked, exactly as video steam, and accepts contracts for media streams from different nodes in the graph. The object developed its own output of the processed data after joining the other pipeline nodes. Using packet objects, which have a broad range of data storage capabilities, each stream of information or data is transmitted to each calculator [8]. It is also feasible to push side packets onto a graph, through which a calculating node may be appended with any further details like constants or static features. The streamlined structure of this pipeline makes it simple to add or change components, and the flow of data is easy to accurately control. The backend machine learning pipeline for the hand tracking system consists of two interrelated models: B is a landmark model,while an is hand detection model [9].The hand detection model delivers the property precisely hand picture to the landmark model. By deploying this technique, deep learning models make use of data augmentation techniques like rotation, flipping,, and scaling allocate more processing power to landmark localization.The traditional approach involves recognising the hand from the frame and then locating landmarks across the current frame. Besides that, in order to tackle ML pipeline concerns, this hand detector takes a different way.The process of recognising and analyzing hands requires thresholding, image processing, and tackling with a variety of hand sizes, all of which take time. The hand detector is trained first, which estimates class labels around objects like fists and palms, which makes it easier to recognise hands with connected fingers than to rapidly identify the hand from the current frame. An encoder-

decoder is also utilized as a selector to extract to make greater scene situations.

3.2.2 Data cleaning and normalization

To aggregate all the sample points into a single file, each picture in the dataset is processed via stage1; comparable to stage 1, we just assess the detector x and y coordinates. Eventually, this file is scraped using a python library named pandas to look for any record with null values. Uncertain images may make it impossible for the detector to recognise the hand which would result in a void entry in the dataset. It is essential to filter these data points because, if not , the prediction model would really be biased. The table's rows containing null objects or entities are identified and eliminated using their indexes. We normalized the x and y coordinates after removing indescribable locations in order to make space for them in the system. After that, the data file is composed to be divided into training and validation sets. The remaining 80% of the data are used to train models using several optimization and loss functions, with the remaining 20% of the data being kept for model validation.

3.2.3. Prediction using Machine Learning Algorithm

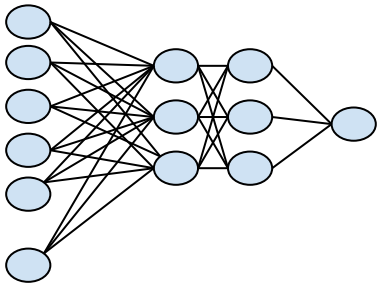

Fig 4: Convolutional Neural Network

Machine learning methods are used to do predictive analysis of several sign languages, and CNN(convolutional neural network) out performed other approaches. Figure 3 depicts the CNN layers and how they operate. As per the diagram, input enters into the input layer first, calculation takes place in the hidden layer, and the computation proceeds to the output layer to produce the result. In high-dimensional spaces, SVM performs well. SVM works effectively when the sample size exceeds the n-dimensional space. SVM is a collection of supervised learning algorithms that can categorize data, conduct regression and find outliers.

3.2.4 Quantitative Analysis

The outcomes for each sample were evaluated using performance metrics as precision, recall, accuracy and F1 score. Accuracy is characterized as the percentage of expected data points that actually happened. ASLR is calculated as the proportion of all accurate forecasts to all measurements in the data.

3.3 Converting Sign Language to Text or Voice

Initially, a person who can't speak and listen provides inputs via hand gestures. The sign will be recognised as a hand movement and inputted as an image. Later evaluation parameters such as a,b,c,d will be the input. Training a model by giving the dataset that will be used to determine how the alphabets rendered the mute person. The model will then recommend a word or phrase that is similar after the alphabets have been recorded and a word has been generated.Afterward, sign gesture recognition will be used to create the phase.The outcome was produced after the phrase was formed.So that normal people hear what silent persons want to say, the text output will eventually be transformed to voice. The primary model is connected to all of the modules. The hand Gesture Recognition was converted to Vocal using CNN algorithm.

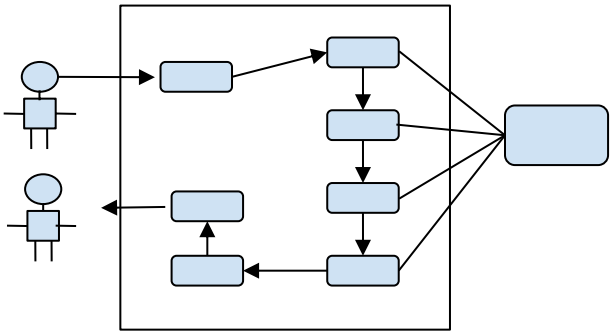

Fig. 5 : Working Procedure of Sign to text

Fig. 6: Sign prediction output

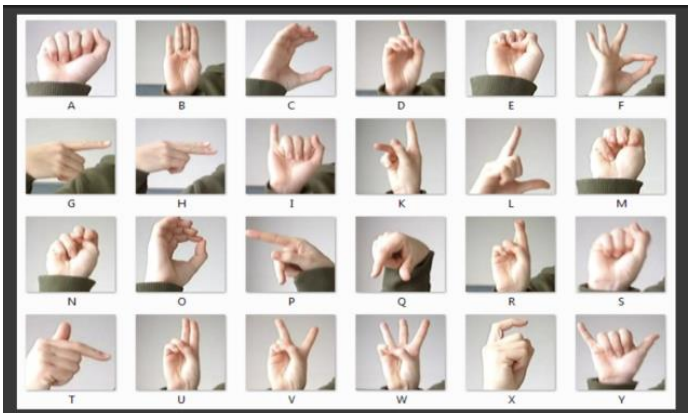

Fig. 7: Dataset for Sign Language

Fig. 8 : Implementation of Sign to text

5. Results and Discussion

To compare our outcome , we assess two metrics. Accuracy of the validation set, or the percentage of properly identified samples, is the most often used metrics in the literature. High accuracy ,essentially measures the probability of classifications where the relevant labeling is included in the Top 5 categories based on score, is another strongly linked matrix. Confusion Matrix was used as well. This enables us to identify the letters that have been misclassified the most and provide recommendations for future improvement.

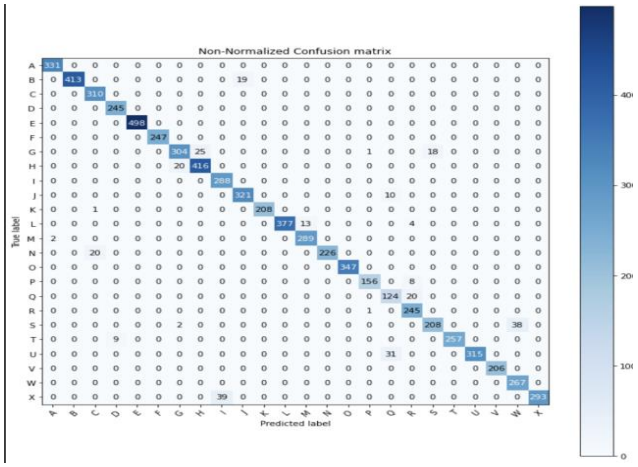

Fig. 9: Confusion matrix of Sign to text

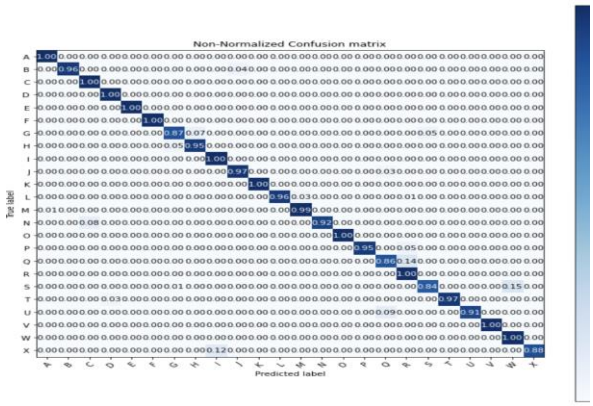

Fig. 10: Confusion Matrix of Sign to text

6. Conclusion and Future Scope

Together utilizing a web application supported by a CNN classifier, we generated and trained an ASL translator. We are capable to develop an effective machine learning model for alphabets a through e and a moderate one for letters a listed in appendix k (excluding).The validation accuracy we achieved during training was not precisely mirrored while testing on the web application because of the lack of variety in database.We recognise that the model will generalized with relatively higher feasibility and and could build a strong model for all words with further data collection in various environmental situations. Additional models we concentrated on GoogLeNet optimisation,however it could be important to look into similar models that have also optimized operations in picture categorizations (i.e VGG and ResNet architecture). Image Preprocessing: we presume that if there is very intensive preprocessing done on the photos, the classification task would be made significantly easier. This could involve cropping, background removal, and contrast enhancement.Using another CNN to identify and trim the hand would be a more reliable strategy. Enhancing the language model : By creating bigram and trigram models would enable us to handle phrases rather than single words. Better letter classification and as well as a more smooth method of retrieving photos from the users at a higher pace are also obligated as a result of this.

Data Availability

A collection of approximately 250 pictures and GIFs with sign language translation of all the alphabets and sentences. As the model required a particular amount of data for training and testing purposes. So the machine is working on 70% data to assist with training. For testing and validation purposes, 30% of the dataset is used. We will expand the dataset because sign language is also a type of language and only having a limited number of words is not enough.

References

[1]Baranwal N, Nandi GC."An efficient gesture based humanoid learning using wavelet descriptor and MFCC techniques. Int J Mach Learn Cybern" 2017

[2] D. Bragg, O. Koller, M. Bellard, L. Berke, P. Boudreault, A. Braffort,N. Caselli, M. Huenerfauth, H. Kacorri, T. Verhoef *et al.*, "Sign language recognition, generation, and translation: An interdisciplinary perspective," *arXiv preprint arXiv:1908.08597*, 2019

[3] Desa, Hazry. "SIGN LANGUAGE INTO VOICE SIGNAL CONVERSION USING HEAD AND HAND GESTURES." 2008

[4] F. Ronchetti, F. Quiroga, C. A. Estribou, L. C. Lanzarini, and A. Rosete,"Lsa64: an argentinian sign language dataset," in *XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016)*., 2016.

[5] G. T. Papadopoulos and P. Daras, "Human action recognition using 3d reconstruction data," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 8, pp. 1807–1823, 2016.

[6] H. Cooper, B. Holt, and R. Bowden, "Sign language recognition," in *Visual Analysis of Humans*. Springer, 2011, pp. 539–562.

[7]Kshitij Bantupalli,Ying Xie,"American Sign Language Recognition using Deep Learning and Computer Vision", IEEE International Conference on Big Data (Big Data),2018

[8] K.S, Tamilselvan & Balakumar, P & Rajalakshmi, B & Roshini, C & S., Suthagar. " Translation of Sign Language for Deaf and Dumb People. International Journal of Recent Technology and Engineering. 8. 2277-3878. 10.35940/ijrte.E6555.018520", 2020

[9]Kurdyumov R, Ho P, Ng J. "Sign language classification using webcam images", pp 1–4, 2011.

[10]Lancaster, Glenn & Alkoby, Karen & Campen, Jeff & Carter, Roymieco & Davidson, Mary & Ethridge, Dan & Furst, Jacob & Hinkle, Damien & Kroll, Bret & Leyesa, Ryan & Loeding, Barbara & Mcdonald, John & Ougouag, Nedjla & Smallwood, Lori & Srinivasan, Prabhakar & Toro, Jorge & Wolfe, Rosalee. "Voice activated display of American Sign Language for airport security.", 2003

[11] O. Koller, S. Zargaran, H. Ney, and R. Bowden, "Deep sign: Enabling robust statistical continuous sign language recognition via hybrid cnn hmms," *International Journal of Computer Vision*, vol. 126, no. 12, pp. 1311–1325, 2018.

[12]P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kauz, "Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural network", in Proc. IEEE Conf. Comput. Vis. Pattern Recog, 2016.

[13]Rekha J, Bhattacharya J, Majumder S. "Hand gesture recognition for sign language: a new hybrid approach. In: International Conference on ImageProcessing, Computer Vision, and Pattern Recognition" (IPCV), pp 80–86, 2011

[14]R. Sharma et al." Recognition of Single Handed Sign Language Gestures using Contour Tracing descriptor. Proceedings of the World Congress on Engineering"Vol. II, WCE 2013, July 3 - 5, 2013, London, U.K.,

[15]R. Sharma, R. Khapra, N. Dahiya. June 2020. Sign Language Gesture Recognition, pp.14-19

[16]Sepp Hochreiter et al.,"Long Short-Term Memory,", Neural Computation 9(8): 1735-1780,1997.

[17]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.

[18]Wang RY, Popović J. 2009. Real-time hand-tracking with a color glove. ACM Trans Graph 28(3):63

[19]Zhang, F., Bazarewsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C. L., & Grundmann, M. 2020. MediaPipe Hands: On-device Real-time Hand Tracking. arXiv preprint arXiv:2006.10214

[20]Z. Ren, J. Yuan, J. Meng, and Z. Zha,"Robust PartBased Hand Gesture Recognition Using Kinect Sensor", " IEEE Trans. Multimedia, vol. 15, no. 5, pp. 1110–1120,2013

