Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 258 (2025) 1980–1992

Procedia
Computer Science

www.elsevier.com/locate/procedia

International Conference on Machine Learning and Data Engineering

Interpretation of Indian Sign Language to Text and Speech to Communicate with Speech and Hearing Impaired Community

Shilpa Ingoley[a]*, Jagdish Bakal[b]

[a]*TSEC, Bandra, Mumbai 400050, India*
[b]*Pillai HOC, Rasayani, Raigadh 410207, India*

Abstract

One of the necessities of a human being is to express themselves. To express one's thoughts, people typically use speech. However, individuals who are deaf and with speech-impairment convey their feelings by practicing sign language. With this communication medium they perform gestures and interpreting these gestures means knowing sign language(SL). Everyone is not acquainted with SL which is the main obstacle in this communication process. To address this key issue, we recommend the deep learning-based approach which assists in interpreting Indian Sign Language(ISL) to text and speech. Communication between a hearing-impaired individual and an illiterate/blind person is made possible through speech conversion. We created a sample model which interprets numbers used in ISL. Numbers are helpful at many places; financial institutions like banks, cash counters, to mention time/distance. Existing sign language recognition systems do not adequately address the challenges posed by i) left, right and ambidextrous signer ii) signers facing mobile camera in selfie-mode iii) signers wearing gloves on their hands to facilitate communication in cold climates or at hill stations iv) If more than one gesture is present for particular sign then those should be considered. To address the mentioned problems, we have created a diverse dataset named ISL_Num14_AS3N24. We have proposed a system wherein we made use of deep learning technique along with our suggested methodology which utilises the hand keypoints in the pre-processing step. Also, for quicker training, we employ modified two pretrained models Vgg16 and MobileNetV2. Text-to-speech conversion module utilizes gtts library. The system exhibits excellence performance with validation accuracy for Vgg16 as 99.76 and for MobileNetv2 as 99.44. Furthermore, the validation loss of VGG16 and MobileNetv2 is 0.9471 and 1.9921 respectively.

Keywords: Deep learning; Hearing-impairment; Speech-impairment; Indian Sign Language(ISL); Pretrained Model; Convolutional Neural N/W.

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: shilpa.ingole@thadomal.org

10.1016/j.procs.2025.04.449

1. Introduction

Speech and hearing-impaired people face a range of significant challenges that affect nearly every aspect of their daily life. Sign language offers solutions but they are not accessible/understood by everyone. Moreover, not all hearing individuals are familiar with sign language(SL), which can create further barriers in day-to-day interactions. These challenges extend beyond mere communication barriers but can impact education, employment, social interactions, and overall quality of their life. We tried to address this issue by using technological aspects. To have smooth interaction with them, SL interpretation is one of the solutions. Besides within the deaf and mute community, sign interpretation is needed at many places in our day to day life. These kinds of systems can be used for aged people who are hard of hearing. Nowadays hearing loss is a major concern and it could happen due to accident or improper habit of using earphones, headphones, mobile etc. Communication in gestures would be helpful in places like hospitals, libraries where we are supposed to maintain silence. Underwater we can't speak, but we can communicate through signs. Similarly, uses of specific sign gestures are common in most of the games, few examples are cricket, hockey, football etc. To manage traffic, traffic personnel use typical sign gestures and the list of applications is almost interminable.

Our work is on numeric hand pose interpretation used in ISL(Indian Sign Language). Numbers are helpful at many places to specify money, time, distance etc. For example: at any cash counters where payments are made, financial institutions like banks, stock broking farms, insurance companies etc., numbers are also helpful in mentioning time and distance which are essential at transportation services like bus stops, railway stations and at airports. We have chosen a numeric 'sign' interpretation that converts sign to text and text to speech. Our system is also considerate about the issues of the existing system. Present SL recognition systems do not adequately address the challenges posed by left hand signer, right hand signer and ambidextrous signers. Signers can belong to any of these categories and in all cases the system should interpret it correctly. Additionally, if signs are captured using the mobile's selfie-mode, which is a mirror image. This may not result in the expected interpretation which can bring down the efficacy of the system. When we try to solve the problem using computer vision, mostly identification is done through appearance or on the basis of coordinate points. Such instances affect the performance of these systems and should be taken into consideration [1]. Therefore, to address the mentioned issues and to have correct interpretation of sign gestures, we have created a diverse dataset which is inclusive of signers' left and right hands gestures. The dataset also consists of images wherein signers' wearing various coloured gloves on their hands to facilitate communication in cold climates or at hill stations. Moreover, in ISL for number 9(nine), two sign gestures are common in practice, our system took this into consideration and accommodated the two different sign gestures for number nine's. We created an adaptable dataset with the assistance from multiple signers both males and females. Signs are captured in diverse backgrounds, varied distance from camera and diminutive variation in orientations. System also has a text to speech conversion module. Speech is a more convenient form of interpretation. Text to speech interpretation has one more benefit, communication between a hearing-impaired individual and an illiterate or blind person is made possible through speech conversion. Our system first interprets numeric signs to text and then text to speech translation is done with the help of 'gtts' library. For the mentioned problems, this research work developed two improved deep-learning models based on the concept of pretrained model with some modification. Beside attaining high accuracy, this approach has also helped in speeding up the training process. To recognise the numeric hand gestures of ISL, we applied the 'keypoints' concept along with our recommended methodology. We worked on two pretrained models namely Vgg16 and MobileNetV2.

The flow of this research work in this paper is as follows: segment 2 is influenced by some of the researchers in the arena of SL. In segment 3, proposed methodology is discoursed in detail along with steps in dataset and model creation. It also comprises a text to speech conversion module. Experimental setup along with implementation details and result analysis is carried out in segment 4. It contains comparative analysis of two different models. Final segment contains the concluding observations.

2. Literature review

This section covers a quick overview of some of the current techniques for sign language recognition used by several researchers. It will discuss some of the contributions made by researchers in ISL and on other sign languages.

Numerous investigations have been carried out and continue to be carried out on gestures, pattern recognition, and character identification in ISL as well as on other sign languages. Different technologies, such as vision-based and sensor-based, also called glove-based/hardware-based methods are utilized in gesture recognition systems to capture hand gestures. Gestures can be static or dynamic. Static gestures are stationary whereas dynamic gestures are generated with movements in hands and/or in body. Vision-based techniques offer greater convenience as users don't need to wear any hardware based gloves or sensors[2]. In vision-based gestures are captured with a camera. A special depth camera is also suggested by researchers for gathering depth data to obtain better precision, however at greater calculation rates. Traditionally, image processing based techniques were utilized for SR recognition. Nevertheless, the accuracy of such systems is generally not high. Hence nowadays researchers have shifted their focus on machine learning and deep learning based solutions. Due to a rise in performance, use of encoder, decoder, transformer etc. are gaining popularity day by day. These types of systems required a significant amount of dataset and high-end computational resources to attain high accuracy. Typically, in glove-based systems, different types of wired or wireless sensors are utilized to identify hand movements and transmit that data to a computer. Although in this kind of system gesture detection is accurate, it comes at a significant cost invested in sensors and wearable devices. Wearing them may cause inconvenience to the users. SL is a complete language where facial expressions are also important; however, sensor based systems cannot identify them, which is a major drawback. There are various sign languages across the globe, every county and region use specific sign language. This is similar to vocal languages, wherein sign gesture, grammar, vocabulary differ region wise. Subsequent section discusses the work related to ISL as well as on different sign languages.

This[3] is a ISL based work which uses gyroscope and accelerometer signals to track the hand position in 3D mode. Input is fed to three DL models for classifications. The MC-DCNN , t-LeNet and modified t-LeNet architecture gives classification accuracies as 83.94%, 79.70% and 81.62% respectively. The [4] describes an alphanumeric ISL system, which recognises users with six and five fingers and translates them into their corresponding equivalent text. They built their custom tailor dataset. They used ML and mediapipe for detection of signs. Overall, they could achieve 90% average accuracy. As SL interpreters are not effortlessly available, the [5] developed a translation system, wherein they translate the text into SL. Their system contains a corpus of English words and sentences. They generate SL based on ISL grammar with the help of 3D avatar animation. This system archives a decent accuracy of 95%. The vision based approach is taken by[6], wherein ISL words are recognized in static, dynamic and with finger spellings. Authors feel the approach proposed by them is user-friendly and can be employed on mobile cameras too. They mentioned due to absence of publicly obtainable dataset on ISL, they created their own dataset. They claim to have developed a signer independent system. For key frame extraction, use of Zernike moments was made which resulted in reducing the computation speed. In fingerspelling alphabets they suggested an improved technique for co-articulation removal. The [7] built a robotic hand based system to recognize ISL gestures, which would be helpful for deaf community. They design the robotic hand interpreter system to interpret different ISL alphabets. The system gives accuracy of 94%. For this the experimental setup of two robotic hands were used, which carried out algorithm. The work proposed by[8] identifies 80 words from a self created SL dataset. Along with the mediapipe they used two models based on SVM and YOLOv4. Their SVM based system is implemented without any supplementary pre-processing and image enrichment operations. SVM attains accuracy of 98.62% whereas YOLOv4 attains 98.8%, which is higher than SVM. They claim that both the systems predict gestures in real time, however SVM is vulnerable to noise whereas YOLOv4 eliminates transition gesture recognition problems with high computation cost. There is no universal SL, hence [9] created an Arabic SL translation(SLT) system. They proposed an SL translation system wherein they used mediapipe along with LSTM and integrated it with a neural network for SLT. They collected their own data for training using a leap motion controller. The pattern-matching algorithm suggested by them judiciously translates the identified gestures into the relevant text. System[10] typically uses computer vision techniques to analyse sign gestures and movements and map them to written or spoken language. SL recognition technology has the potential to greatly improve the accessibility of communication for people with hearing and speech impairments and to improve communication between people who speak different languages. In this paper, the projected system has attained the accuracy of 91.67% which is better associated with the existing works in the literature. Authors[11] have an opinion which says that due to the lack of publicly available SL datasets on South African SL, there is dearth of work in the SLR area. They applied a Deep-CNN based model using 26 static African SL alphabets. Data preprocessing and augmentations techniques were used in implementation. They also made use of google APIs to translate sign output to different south African languages. Their model outperforms with 98% weighted average accuracy. To identify the basic healthcare domain related words in real time using a webcam, [12] proposed a system. They compared three methods and concluded that their proposed method based on OpenPose and LSTM is the most effective one among all others and attains approximate accuracy of 97%. They used a self-created dataset with 20 classes and also created three systems namely: CNN+ ImageStacking, CNN+LSTM and LSTM+OpenPose. Author [13] demonstrates SL recognition taxonomy, wherein input sources such as camera, wearable devices and datasets were discussed. They also discussed preprocessing and feature extraction methods. For ASL alphabet recognition, implementations are performed using CNN based and InceptionV3 algorithm. The favourable accuracy of 98.94% is attained with the InceptionV3 algorithm whereas CNN based exhibits 98.54% accuracy. SL interpretation done by[14] uses facial expression along with hand gestures by introducing the method DFCNet+ (Dynamic Feature Contrast Net Plus). It incorporates both cross-model learning and dynamic feature extraction. Validation of method is done on benchmark datasets: PHOENIX14/-T(German) and CSL-Daily(Chinese). As per[15] many SL categories exist as there are more than 300 sign languages present worldwide. Two main challenges of SLR are: real world signers may not be present in the dataset. Second, creating a large scale dataset is a labour-intensive and time-consuming process. To address these challenges they construct a signer-independent learning SLR method for the single-frequency dataset and achieve the best result on the CSL-500 dataset. The hardware and cloud based approach used by[16], proposes a TinyML(Tiny machine learning) solution for SL recognition system using a wearable, low-cost, IoT(internet-of things) device. An edge device to interpret isolated signs from the ISL, by using the time-series-data collected from the motion-sensors of the device. The system achieves an average 87.18% accuracy. SignTalk was devised for text-to-speech conversion.[17] believe that establishing a bidirectional system will benefit the deaf community's mental health as well as job prospects. Hence, they offer solutions to facilitate SL to text/speech and vice versa. Input to their system can be given by both methods: data-glove(hardware-based) or camera based. They use an avatar for sign language performance. They have done word and character level ASL translation. [18] applied a creative approach which has linguistic modelling of the corresponding text of SL during the process of converting sign to GLOSS. They used text correction module in the implementation: first predictor, corrector and final predictor. To evaluate effectiveness of their framework they worked on RWTHPHOENIX-Weather-2014-T and CSL dataset and results showed enhanced accuracy of the approach.

Table 1. Comparison of different SL Interpretation

Reference/Year	SL	Technique used	Dataset(s)	Accuracy	Sign Identification	Images /Videos
[22] / 2022	ISL	SURF with SVM and CNN uses the model: Bag of Visual Words (BOVW)	custom-built dataset	SVM: 99.17% CNN: 99.64%	Identifies ISL letters (A-Z) and digits (0–9)	Videos
[23] / 2020	Hand Posture; ASL	Deep CNN architecture	-NUS hand posture -American fingerspelling A dataset	NUS: 94.7 ± 0.8 % ASL: 99.96 ± 0.04 %	-10 handPosture -24 ASL alphabet	Images
[24] / 2024	Chinese SL(CSL) Arabic SL	Spatiotemporal feature-based method CNN TD(Time Distributed)	-RGB Arabic Alphabet Sign Language (ArSL) -CSL dataset	CSL : 90.87% ArSL : 89.46%	32 alphabets for both CSL and ArSL	video
[25] / 2021	ISL	Graph matching concept; 3D motion capturing technology - EEMGM	3D ISL dataset created with 8 IR and 1 RGB cameras	97.38%	350 ISL words	3D videos
[26] / 2024	Bagla SL	pre-trained CNN models	Own developed: ‘‘BdSL_OPSA22_STATIC1’’;‘‘BdSL_OPSA22_STATIC2’’	98.38% : static1 92.78%: static2	46: Numerals Alphabet	images
[27] / 2023	-ASL -Bagla SL	Pre-trained YOLOv5 with SE and CBAM attention modules.	ASL and OkkhorNama	98.9 %: ASL 97.6 %: BdSL	alphabetic and numeric 36: ASL	images
[28] / 2022	ISL	LSTM, LSTMBi, LSTM stack, mediapipe	synthetic dataset created	100% to 92.68%	seven gestures; 26 alphabets	videos

[29] / 2021	ISL	LSTM, LSTM Seq2Seq, transformer	created a dataset	BLEU score for test data 0.6751.	55 different sentences	videos
[30] / 2020	ISL	CNN, LSTM, RNN	self-recorded ISL dataset	Bank:85%, Everyday: 90%, Entire dataset:81%	20 signs	video
[31] / 2023	ASL	CNN	ASL Kaggle Dataset	96.3%	26 alphabets	Images
[32] / 2020	ISL	Deep neural networks, Encoding and Decoding.	Created: INCLUDE INCLUDE-50	85.6% INCLUDE 94.5% INCLUDE50	263 words 50 words	Videos
[33] / 2023	Assamese SL	Feed forward neural network, Mediapipe	Created using webcam and Microsoft Kinect	99%	9 (Vowels and consonants)	2D/ 3D images,
[34] / 2021	ISL	KNN, SVM, CNN	self-created	KNN: 98.3; SVM: 98.7; CNN: 99.1	36 (alphanumeric)	images

A paper[2] is reviewed on automatic SL recognition. Total publications studied were 649 related to SLR from the Scopus database. They analysed various good results giving features extraction techniques and classification techniques. They had performed comparisons and mentioned in detail about different techniques on data/image acquisition, enhancement, preprocessing and many more. According to them, the perfect SLR system till today is an open problem. The [19] is a survey on SL literature. The paper provides a wide-ranging review of pertinent research carried out in this area. They have multiple viewpoints on SL recognition and translation systems. They emphasise that SL is a complete language hence hand movement, facial expression and body posture integrated recognition should be carried out at a time. They highlight the major challenge in SL recognition is availability of benchmark datasets. To address the issue of hearing and hearing-impaired communities[20] it's challenging to build a two-way communication system. They discuss the challenges, recent developments, fundamental components required in a bi-directional SL translation system, advantages, limitations along with possible lines for future research. They have mentioned the backbone architecture for this is CNNs, Transformer, Generative models, Motion capture and signing avatars etc. For virtual engagement in these digital timings, [21] emphasised the real time SL translation is needed. Systems like TEAMS(Microsoft), Zoom are commonly used online platforms. In real time besides flawless recognition, excellent responsiveness and low latency is desirable. They have provided a guideline in this assertive-technology for further research. Graphically, their work mentioned the frequency of various datasets used in SLR. Table 1 illustrates a comparison of several parameters applied by researchers in their work.

3. Methodology

The proposed methodology has been separated into three main subsections: Dataset creation, Model creation and Speech conversion. Fig. 1 shows the major steps in creating overall system architecture. It indicates that after capturing the sign gesture from the camera is a raw image. For identification of presence of hand and landmarks on it, mediapipe has been used[8]. The library Mediapipe is an open-source project created by Google. It offers a range of modules, from which we utilize two models: hand tracking and palm detection. Images are captured using OpenCV. BGR is the default colour format in OpenCV, hence needing to be changed to RGB before feeding this as an input. Presence of hand(s) detection is done by palm detection model. If it is detected, then it identifies the Hand Landmark, these are 21 key landmark points on the hand(s). We want these landmark keypoint on the image, we are not interested in collecting the coordinate positions of the ‘x’, ‘y’ and ‘z’ axis. After identification of these points on hand(s), we create a bounding box around it without touching the outline of the hand/fingers. The Bounding box comprising the hand image is cropped from the entire image. As every number sign gesture takes different shapes; copped images are not the same in height and width. Hence needs to add padding to make them the same in height and width. For padding any colour can be used, we applied white. If the cropped image is vertical or horizontal then padding is applied on the image, which can be seen in Fig. 1 module second preprocessing steps. These images are fed to the model for training and classification. Predicted numerals of each class are mapped to text. After finding correlation in adjacent images, it will be converted into appropriate text and finally given to a speech module for audio generation. For text to voice conversion, we used the ‘gtts’ library, it will convert text to .wav file to speech.

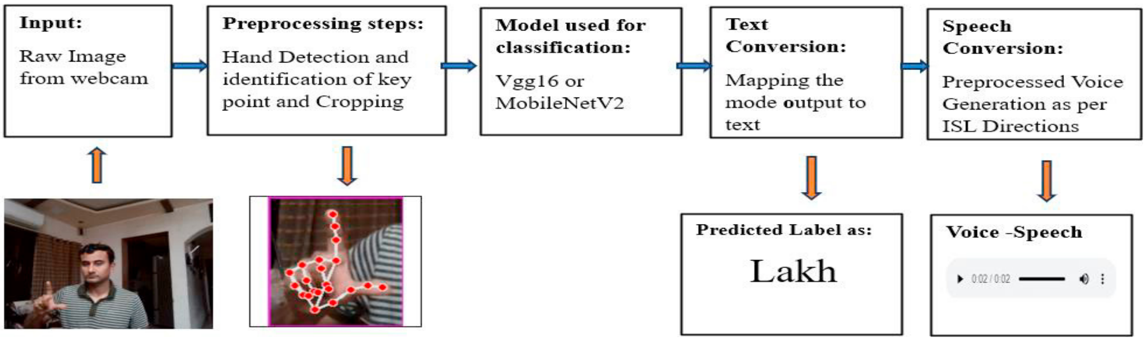

Input:
Raw Image from webcam
Preprocessing steps:
Hand Detection and identification of key point and Cropping
Model used for classification:
Vgg16 or MobileNetV2
Text Conversion:
Mapping the mode output to text
Speech Conversion:
Preprocessed Voice Generation as per ISL Directions
Predicted Label as:
Lakh
Voice -Speech
0:02 / 0:02

Fig. 1. Major Steps in system creation/architecture

3.1. Dataset Creation

For the mentioned purposed we have developed Dataset name: ISL_Num14_AS3N24, where ISL stands for Indian Sign Language, Num14 is for fourteen numeric signs and AS3N24 indicates the name of the signers: Ashish, Sarwasvi, Swaroop, Shilpa and Nitin; followed by 24 which is the dataset collection year 2024. For dataset creation following steps are followed: To capture ISL numerals, no use of any special camera or depth camera were made for image capturing, simply web camera of laptop was utilized. Signers were asked to perform different numeric gestures facing the camera. To capture images, OpenCV is used. Sign gestures performed by five signers both male, female from different age gaps ranging from teenage to fifties. Also, one of the signers wearing different coloured gloves performed signs. No constraints were applied on the clothing of signers. To make the dataset robust and versatile, Fig. 2, shows signers making different hand postures in varied backgrounds, variation in angles and distance from the camera. If there are multiple sign gestures used for a particular sign, they should be taken into account for accurate sign interpretation[35]. Hence two common gestures of number 9 have been considered. The dataset consists of 14 classes, namely 0 to 9, 9' (9' represents the alternative gesture for number nine), Thousand, Lakh and Crore. For each class a total 450 sign gestures are considered. A total of 6300 images are obtained without any augmentation technique. To create a randomness, specific code is given to each user so that the resulting outcome is the shuffled images. All the images in the dataset are supplied to Mediapipe's hand detection model to produce keypoints on hand(s). Without touching the silhouette of the hand, a bounding box is created to crop the hand gesture from the image. All cropped images will not be of the same size. To convert it into the same size, the concept of padding is used. Thus, cropped images are converted into equal height and width and are stored in respective labelled sign gesture folders. Consequently, the dataset is created for all the images and can be utilized for model creation.

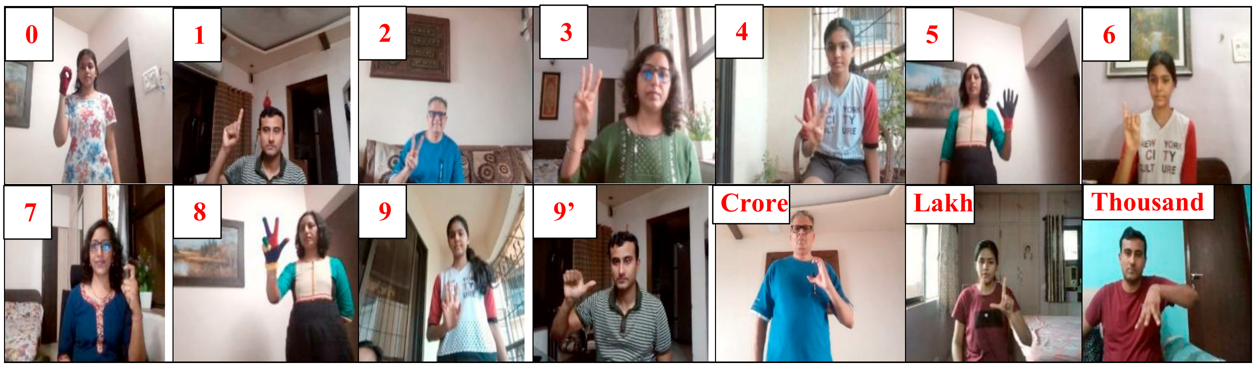


Fig. 2. Considered sample sign for ISL numeric, some of the pictures shows signers are wearing glove

3.2. Model Creation

The following section describes the model creation procedure adopted by us to create two models namely Vgg16 and MobileNetV2 indicated in Fig. 3, from dataset ISL_Num14_AS3N24. Before feeding the images to models, images are reshaped to 224*224*3, where 3 is for RGB channels; no grey scale conversion is done, however normalization on images is performed. Normalization is the procedure of scaling the pixel values of an image to a fixed typical value between 0 and 1. Dataset built is inherently versatile; hence no data augmentation technique was utilized. The Models are used for feature extraction, where the concept of transfer learning(TL) is leveraged. This facilitates retaining the pretrained weights, which would be helpful for ISL numerals classification. The final layer in both the models is the 'softmax' layer, which gives the classification probability for the respective classes. In this case, the number of neurons is 14, which is equivalent to the ISL numeric classes considered.

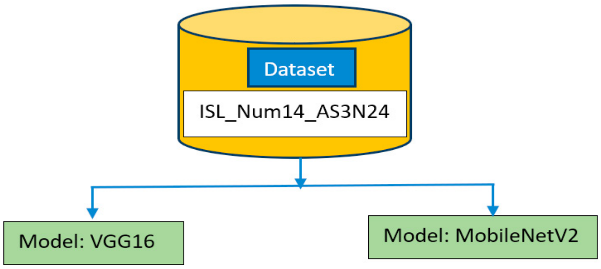


Fig 3: Two models created using dataset: ISL_Num14_AS3N24

3.2.1. Model VGG16

Vgg16 or 'VGGNet' is a deep 16-layers CNN model architecture that was developed by the VGG(Visual Geometry Group) at the University of Oxford. Sixteen weight layers are present in the network, which includes 13 convolutional layers and 3 FC(fully connected) layers. Convolutional layers use small 3*3 filters with a stride of 1 and 2*2 max pooling layers with a stride of 2 to reduce spatial dimensions. In Fig. 4, the first block indicates a based model, which is already pre-trained on 'ImageNet' dataset and its weights are frozen. This is the transfer learning concept wherein weights are blocked so as to retain the learnt features on earlier dataset. This helps in reducing the training time and hence computational resources. Next step is training the subsequent block, wherein we trained the last three fully connected layers by modifying the neuron sizes to 256, 128 and 14 respectively.

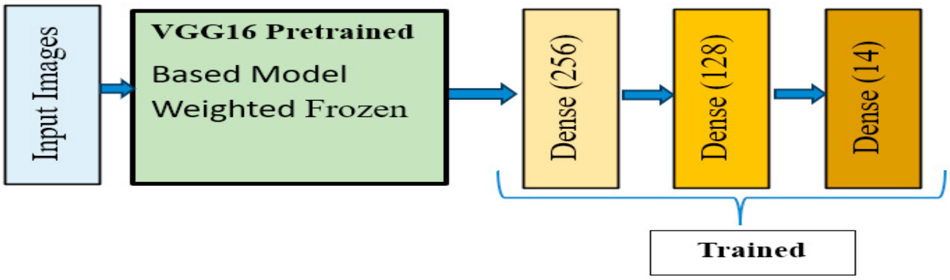


Fig. 4. VGG16 implemented architecture

3.2.2. Model MobileNetV2

MobileNetV2 is a CNN designed specifically for mobile and edge devices by Google Research. MobileNetV2 is notable for its ability to deliver high performance with relatively low computational resources, making it well-suited for mobile and embedded applications. MobileNetV2 presents the concept of inverted residuals, which use lightweight depthwise, pointwise separable convolutions to decrease the number of parameters and computational cost while maintaining accuracy. The construction of model mobilenetv2 is shown in Fig. 5. The pretrained model is trained on ImageNet dataset and we used weights to leverage learned features for identification of ISL numerals. The fine-tuning on our self-created datasets were performed. Subsequent block to base model described the modified architecture. Series of layers appended are: average Pooling2D, Flatten, Dense layer with size 256, Batch normalization, Dropout, Batch normalization and to end a FC Dense layer with size 14 used for classifying ISL numerals.

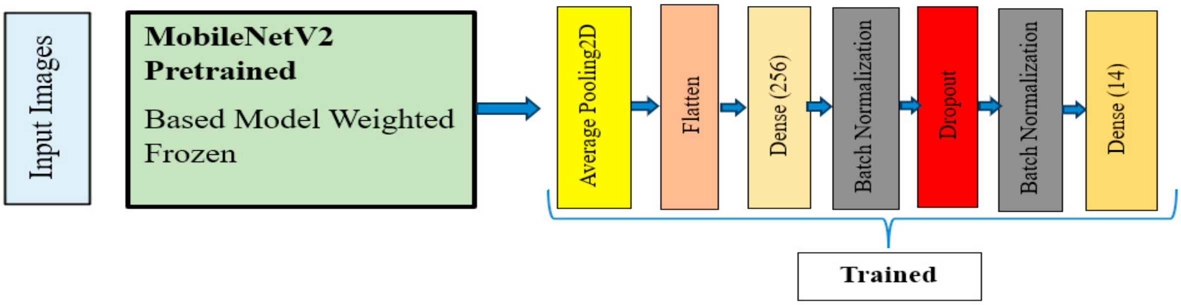


Fig. 5. MobileNetV2 implemented architecture

3.3. Speech Conversion

To convert text to speech, we have utilized gTTs(Google Text-to-Speech) , a Python library. To translate the predicted output into appropriate text certain things we need to take into consideration before converting it into audio. First, we are required to find the correlation in adjacent images, then it will get converted into appropriate text. For example, to sign the number hundred (100), the signer is going to perform three sign gestures namely one, zero and zero. These three signs mean a hundred. Likewise, we have considered two gestures of nine, these two different gestures mean the same number nine. To generate audio, we selected English with a British accent. The generated speech is saved in MP3 as an audio file, which can then be played back.

4. Implementation and Results Analysis

The proposed methodology is coded in Python and trained on Google's 'colab', with the following details: Type of runtime: Python 3. Hardware accelerator: T4 GPU, Tensorflow version: 2.15. To ensure a fair comparison, both models receive the same set of parameters. Table 2 displays the details of parameter settings at implementations.

Table 2. Parameters for Vgg16 and MobileNetV2

Parameter	Values
Number of classes	14
Training sample per class	360
Validation sample for each class	90
Split ratio (training: validation)	80:20
Optimizer	Adam
Learning rate	0.001
Loss function	categorical_crossentropy
Pretrained Models used	Vgg16 and MobilenetV2
Batch Size	64
Early Stopping technique applied	Yes

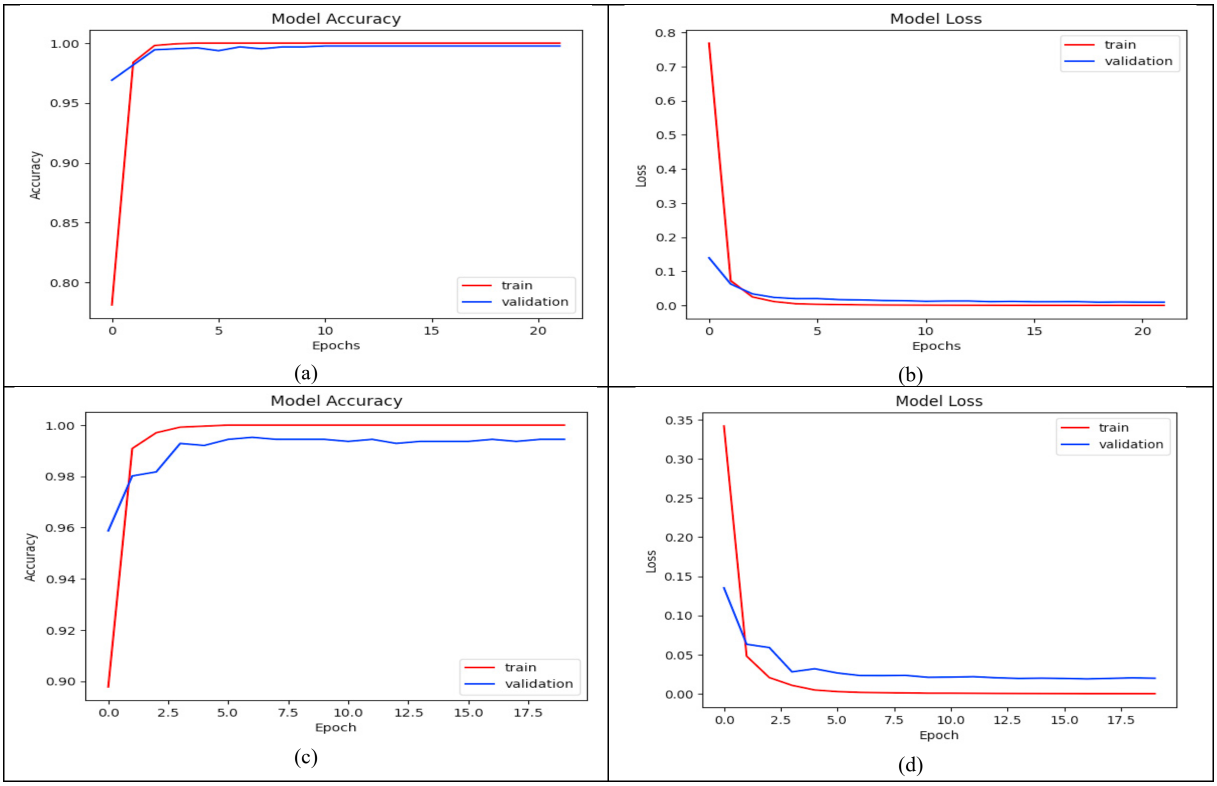
The suggested technique has been trained and validated on ISL Numeral dataset: ISL_Num14_AS3N24. We created two models: Vgg16 and MobileNetV2 that can be observed in Fig.4 and Fig.5 respectively. Dataset is split in the ratio of 80:20, for training 80% images are utilized from each class and the remaining 20% is reserved for validation. For both the models input image size is resized to 224*224*3, as each original model was verified with this image size, hence to enhance the outcome we kept it the same. This is a multiclass classification problem, where numerals are identified hence the activation function used in the last layer is ‘softmax’ and with value 14 is passed. Equation 1 illustrate the formula to compute the softmax function, which gives the classification probability of each class.

$$y_j = \frac{e^{x^T w_j}}{\sum_{k=1}^{K} e^{x^T w_k}} \tag{1}$$

Where standard exponential function is represented by $e^{x^T w_j}$, for input vector. The multiclass classifier 'k' represents the number of classes. For output vector, $e^{x^T w_k}$ is standard exponential function. The class with the highest probability would be the output. To measure the losses "categorical_crossentropy" function has been used and an equation for the same is shown in equation 2.

$$categorical\_crossentropy\_loss = -log\frac{e^{s_p}}{\sum_j^C e^{s_j}} \tag{2}$$

Where the net scores in classes C is s_j, s_p represent the CNN score for the positive class. We have well-thought-out on 'adam' optimizer in our implementation, due to its quick conversion rate and computational efficiency. To stop the training process and to avoid models getting overfit, we have applied an early stopping technique on both the models, wherein 'validation losses' are monitored. Though the number of maximum epoch values is assigned to 100, the patience value is set to 3 to terminate the training. Fig. 6. illustrates the graphs for accuracies and losses. Fig. 6(a) and Fig. 6(b) demonstrate the accuracy and loss curves along with the number of epochs for Vgg16 model. Similarly, Fig. 6(c) and Fig. 6(d) demonstrate the accuracy and loss curves for Model MobileNetv2 against the number of epochs.



(a)

(b)

(c)

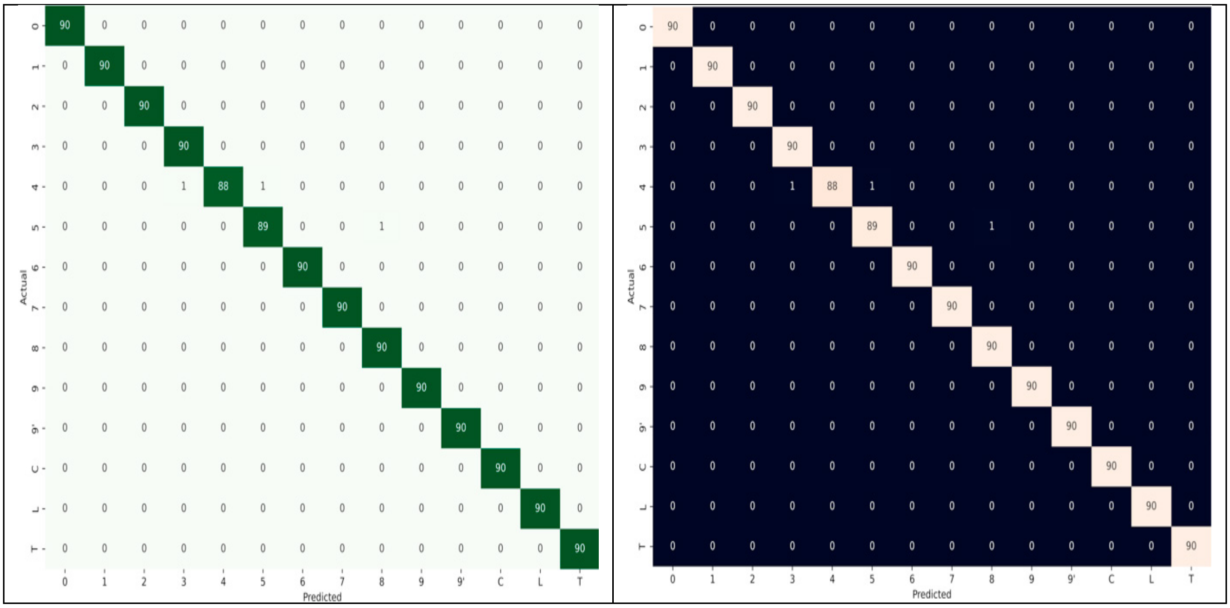
(d)

Fig. 6. Graphs of accuracies and losses for models (a) Vgg16 model Accuracy vs Epochs (b) Vgg16 model Loss vs Epochs (c) MobileNetV2 model Accuracy vs Epochs (d) MobileNetV2 model Loss vs Epochs

Table 3. Average values for precision, recall, F1 score and support for models Vgg16 and MobileNetv2

Model	Precision	Recall	F1-score	Support
Vgg16	0.997	0.997	0.995	90
MobileNetV2	0.995	0.995	0.993	90

Confusion matrices for models Vgg16 and MobileNetV2 are displayed in Fig. 7(a) and Fig. 7(b) respectively. This tabular layout helps in visualizing the outcome of predicted vs actuals. From these evaluation matrices, we can figure out very less misclassification is performed by both the models, nevertheless to mention Vgg16 is better. In confusion matrices, '*C*' stands for Crore, '*L*' is for Lakh and '*T*' is for Thousand. Table 3 indicates the regularly used different performance measurement parameters for evaluation metrics, it shows the average values computed for precision, recall, f1-score and support for models Vgg16 and MobileNetV2. It also designates the use of a balanced dataset.



	0	1	2	3	4	5	6	7	8	9	9'	C	L	T
0	90	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	90	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	90	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	90	0	0	0	0	0	0	0	0	0	0
4	0	0	0	1	88	1	0	0	0	0	0	0	0	0
5	0	0	0	0	0	89	0	0	1	0	0	0	0	0
6	0	0	0	0	0	0	90	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	90	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	90	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	90	0	0	0	0
9'	0	0	0	0	0	0	0	0	0	0	90	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	90	0	0
L	0	0	0	0	0	0	0	0	0	0	0	0	90	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	90

	0	1	2	3	4	5	6	7	8	9	9'	C	L	T
0	90	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	90	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	90	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	90	0	0	0	0	0	0	0	0	0	0
4	0	0	0	1	88	1	0	0	0	0	0	0	0	0
5	0	0	0	0	0	89	0	0	1	0	0	0	0	0
6	0	0	0	0	0	0	90	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	90	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	90	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	90	0	0	0	0
9'	0	0	0	0	0	0	0	0	0	0	90	0	0	0
C	0	0	0	0	0	0	0	0	0	0	0	90	0	0
L	0	0	0	0	0	0	0	0	0	0	0	0	90	0
T	0	0	0	0	0	0	0	0	0	0	0	0	0	90

Fig. 7. Confusion matrix for (a) Vgg16 Model (b) MobileNetv2 Model

Table 4. Comparison between Models Vgg16 and MobileNetv2

Models	Training Accuracy	Validation Accuracy	Training Loss	Validation Loss	Number of epochs taken	Size (MB)	Time taken for training (Minutes)
Vgg16	100.0	99.7619	0.0135	0.9471	22	130.1	11.8421
MobileNetv2	99.9404	99.4444	1.3939	1.9921	20	42.8	6.1057

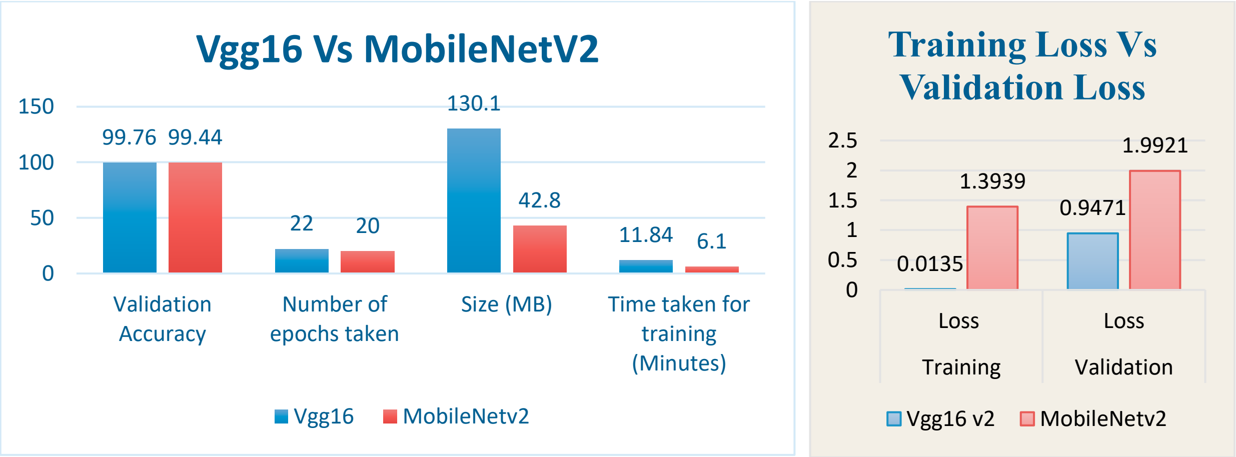


Fig. 8. (a)Comparison between Vgg16 Vs MobileNetV2 (b) Training Loss Vs Validation Loss

Table 4 indicates comparison made on various parameters amongst both the models. Models derive features very well in spite of the very small inter class variations, it discriminates the hand postures accurately. If we evaluate the performance of both the models, it clearly indicates training and validation accuracy of vgg16 is high which is 100 and 99.7619 when compared with MobileNetv2 99.94 and 99.44 respectively. Similarly, losses of model Vgg16 are less when compared to MobileNetv2. However, the number of epochs taken to complete the training, the size of the model(considered in MB) and time taken for execution(evaluated in minutes) are noticeably less in case of MobileNetv2 model. Both the models have attained great validation accuracy of 99.7619 and 99.4444 respectively. Accuracy of Vgg16 is higher as compared to MobileNetv2, however it took 22 epoch and 11.8421 minutes to get trained whereas model MobileNetv2 is 3 times lighter than Vgg16 and took 20 epochs for training and completed in 6.10 minutes. Same is illustrated with the help of a graph and can be observed in Fig. 8.(a). Training vs Validation losses for both the models can be observed in Fig. 8(b). It clearly indicates that vgg16 has learnt the features in a better manner when compared to MobileNetv2. Hence, we recommend to install model vgg16 when resource constraints are not an issue. MobileNetv2 is a lightweight computational model that can be used on resource constrained devices such as mobile and edge. Table 5 gives comparative details of SL implementation by other researchers with our proposed approach. It comprises technique(s) and specification of dataset(s) used along with the achieved accuracy.

Table 5. Comparison of various SL recognition approaches

Ref/ year	**Technique used**	**Dataset Description /subjects/ number of images**	**Accuracy of system(s)**	**Signs Identified**	**Conversion to text /speech**
[17] /2024	Spatio-temporal features + Time distributed CNN	ArSL (Arabic) with 7,857 images CSL (Chines) with a total of 360 videos	ArSL: 89.46 % CSL : 90.87 %	Arabic and Chines Alphabet	Text
[36] /2023	(YOLOv4) (SVM) with media-pipe.	self-created with a total of 676 images.	YOLOv4 : 98.8% SVM+ media-pipe: 98.62%	80 sign words	Text
[37] /2023	Naive Bayes, SVM, RF, KNN	Build dataset with 6 signers and 4121 data points	NB: 80.34% RF: 95.75% KNN: 96.72% SVM : 97.20%	ASL alphabets (26)	Text
[38] /2024	CNN ResNet-50	LSA64 dataset comprises 3200 videos performed by 10 subjects	97.5%	64 different words	Text
[39] /2024	LSTM	Created 30 frames for each sign gesture	Alphabet: 87.50 Dynamic Sign: 97.20 Emergency Sign: 88.88	ISL alphabets (26); 6 dynamic signs, 6 emergency signs	Text
Our Proposed approach	Mediapipe + Vgg16, MobileNetV2	Created with 5 Signers + 1 with gloves ; With total images 6300	Vgg16: 99.7619 MobileNetv2: 99.4444	ISL numerals (0 to 9,9', thousand, Lakh, core) signs To Text and speech	Text and Speech

5. Conclusion

Individuals with hearing and speech-impairment convey their feelings by practicing sign language. Through this communication medium they perform gestures and interpreting these gestures means knowing sign language(SL). Everyone is not acquainted with it and it creates obstacles in the communication process. This research provides interpretation of ISL numeric using deep learning techniques along with our suggested methodology. Numbers are helpful at many places; to specify amount, time, distance etc. One of the applications of our system is similar to PayTM. (In India, many shops and stores have a sound system that announces the amount received when a transaction is completed.) Hence, correct interpretation is vital. In ISL for number 9(nine), two sign gestures are common in practices, our system takes this into consideration. Nowadays mobile is ubiquitous, where cameras can be faced in normal or selfie-mode. To have correct interpretation in this mode and also accommodate left, right and ambidextrous signers we have created datasets using both left and right hands of signers. To facilitate communication in cold climates or at hill stations, we have captured some sign images wherein signers wearing different coloured gloves. Dataset is prepared with the support from multiple signers, both male and female from different age gaps. The dataset is named as ISL_Num14_AS3N24. System is trained using two pretrained models with customization done on top layers of the model to take the benefit of TL(transfer learning). It converts sign gestures to text and then to speech. Two models were created using Vgg16 and MobileNetv2. Both the models have attained great validation accuracy of 99.76 and 99.44 respectively. Accuracy of Vgg16 is higher as compared to MobileNetv2. However, it took 22 epochs and 11.84 minutes to get trained whereas MobileNetv2 is 3 times lighter than vgg16 and took 20 epochs for training and completed training in 6.10 minutes. Training and validation losses for Vgg16 are less in comparison to MobileNetv2 which is the indication that the model vgg16 has learnt the features well. Because of vgg16's higher accuracy, we advise installing and utilizing the VGG16 model. However, Mobilenetv2 has a lightweight design, hence for resource-efficient operation, the Mobilenetv2 model is recommended on devices such as smartphones, tablets, etc.

For future direction the main aim is to develop an effective numeric ISL recognition system. This will be apt for real-world applications to facilitate communication with the deaf-mute community. Furthermore, the goal is to make the complete numeric ISL bidirectional translation system: ISL numbers to text/speech and numeric speech/text to ISL gesture. Building the complete ISL interpretation system which identifies both hand gestures with facial expressions would be challenging in terms of resource requirement and availability of datasets. Hence applying proposed methodology as a sub system can be beneficial to recognize signs performed as a supportive system in facial gestures recognition. Methodology can be employed to develop some of the light weight systems which are essentially needed by deaf-mute community at various common transportation places like at bus stops, railways station, airport etc. and public places like hospitals, banks, government/private offices and so on.

References

[1] Ingoley, S. N., & Bakal, J. W. (2023). "Use of Key Points and Transfer Learning Techniques in Recognition of Handedness Indian Sign Language." *IJRITCC*, 11, 535–545. https://doi.org/10.17762/ijritcc.v11i9s.7465.

[2] Adeyanju, I. A., Bello, O. O., & Adegboye, M. A. (2021). "Machine learning methods for sign language recognition: A critical review and analysis." *Intelligent Systems with Applications*, 12, 56. https://doi.org/10.1016/j.iswa.2021.20

[3] Rinki Guptaa, Sreeraman Rajan.(2020). "Comparative Analysis of Convolution Neural Network Models for Continuous Indian Sign Language Classification." *CoCoNet'19, Procedia Computer Science* 171 (2020) 1542–1550

[4] L. Dias, K. Keluskar, A. Dixit, K. Doshi, M. Mukherjee, and J. Gomes (2022)"SignEnd: An Indian Sign Language Assistant," *IEEE Region 10 Symposium, TENSYMP 2022, Institute of Electrical and Electronics Engineers Inc., 2022*. doi: 10.1109/TENSYMP54529.2022.9864359.

[5] Sugandhi, Parteek Kumar, and Sanmeet Kaur. (2020). "Sign Language Generation System Based on Indian Sign Language Grammar." *ACM Trans. Asian Low-Resour. Lang. Inf. Process*. 19, 4, Article 54 (April 2020), 26 pages. https://doi.org/10.1145/3384202

[6] P.K. Athira, C.J. Sruthi, A. Lijiya.(2022). "A Signer Independent Sign Language Recognition with Co-articulation Elimination from Live Videos: An Indian Scenario." *Journal of King Saud University – Computer and Information Sciences* 34 (2022) 771–781

[7] Yash Verma, R.S Anand. (2023). "Design and control of a robotic hand for generating gestures based on Indian Sign Language*." ELEXCOM)* | 979-8-3503-0511-1/23 IEEE | DOI: 10.1109/ELEXCOM58812.2023.10370694

[8] R Sreemathy , MP Turuk, S Chaudhary, K Lavate, A Ushire, S Khurana. (2023). "Continuous word level sign language recognition using an expert system based on machine learning." *International Journal of Cognitive Computing in Engineering* 4 (2023) 170–178

[9] Abdalla, A., Alsereidi, A., Alyammahi, N., Qehaizel, F. B., Ignatious, H. A., & El-Sayed, H. (2023). "An Innovative Arabic Text Sign Language Translator." *Procedia Computer Science*, 224, 425–430. https://doi.org/10.1016/j.procs.2023.09.059

[10] Hridoy Adhikari, Md. Sakib Bin Jahangir, Israt Jahan, Md. Solaiman Mia, Md. Riad Hassan.(2023). "A Sign Language Recognition System for Helping Disabled People." *2023 5th International Conference on Sustainable Technologies for Industry 5.0 (STI)* | 979-8-3503-9431-3/ 2023 IEEE | DOI: 10.1109/STI59863.2023.10465011

[11] Tebatso Gorgina Moape, Absolom Muzambi, Bester Chimbo. (2024). "Convolutional Neural Network Approach for South African Sign Language Recognition and Translation." *ICTAS* | 979-8-3503-1491-5/24/IEEE | DOI: 10.1109/ICTAS59620.2024.10507130

[12] Jash Gandhi, Parth Gandhi, Aayush Gosar, Sheetal Chaudhari. (2021). "Video Recognition Techniques for Indian Sign Language in Healthcare Domain." *INCET* | 978-1-7281-7029-9/20/ IEEE | DOI: 10.1109/INCET51464.2021.9456116

[13] Shruti Kankariya, Kanak Thakre, Urvi Solanki, Sneha Mali, Abhishek Chunawale. (2024). "Sign Language Gestures Recognition using CNN and Inception v3." *ESCI* | 979-8-3503-0661-3/24/IEEE | DOI: 10.1109/ESCI59607.2024.10497401

[14] Yuan Feng, Nuoyi Chen, Yumeng Wu, Caoyu Jiang, Sheng Liu, Shengyong Chen. (2024). "DFCNet+: Cross-modal dynamic feature contrast net for continuous sign language recognition", *Image and Vision Computing 151* (2024) 105260

[15] Tianyu Liu , Tangfei Tao * , Yizhe Zhao , Min Li , Jieli Zhu.(2024). “A signer-independent sign language recognition method for the single-frequency dataset.” *Neurocomputing* 582 (2024) 127479

[16] S. Sharma, R. Gupta, A. Kumar. (2024). “A TinyML solution for an IoT-based communication device for hearing impaired.” *Expert Systems With Applications* 246 (2024) 123147.

[17] Amira Elnashara, Karim Hamdana , Sultan Al Seiaria, Yusuf Shanableha , Gerassimos.(2024). “Bi-directional translation methods of ASL and speech/text”, *Procedia Computer Science* 239 (2024) 1879–1886

[18] Mohd Faisal, Angad Singh, Dr. Shailendra Singh. (2024). “A Review of Real-Time Sign Language Recognition for Virtual Interaction on Meeting Platforms.” *CONFLUENCE 2024*, 979-8-3503-4483-7, IEEE, DOI: 10.1109/CONFLUENCE60223.2024.10463439.

[19] Alaghband, M., Maghroor, H. R., & Garibay, I. (2023). “A survey on sign language literature.” *Machine Learning with Applications,* 14, 100504. https://doi.org/10.1016/j.mlwa.2023.100504

[20] Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, Vassilis Athitsos, Mohammad Sabokrou. (2024). “A survey on recent advances in Sign Language Production”, *Expert Systems With Applications 243 (2024*) 122846

[21] Xuebin Xu , Jun Fu. (2024). “A two-stage sign language recognition method focusing on the semantic features of label text.” CSI International Symposium on AISP, IEEE | DOI: 10.1109/AISP61396.2024.10475205.

[22] Katoch, S., Singh, V., & Tiwary, U. S. (2022). “Indian Sign Language recognition system using SURF with SVM and CNN.” *Array, 14. https://doi.org/10.1016/j.array.2022.100141*

[23] Adithya, V., & Rajesh, R. (2020). “A Deep Convolutional Neural Network Approach for Static Hand Gesture Recognition.” *Procedia Computer Science, 171, 2353–2361.* https://doi.org/10.1016/j.procs.2020.04.255

[24] Renjith Sa,Manazhy Rashmib,Sumi Suresh M.S.(2024). “Sign Language Recognition by using Spatio-Temporal Features.” *5th International Conference on Innovative Data Communication Technologies and Application(ICIDCA 2024). Procedia Computer Science* 233(2024)353-362

[25] E. Kiran Kumar, P.V.V. Kishore , D. Anil Kumar, M. Teja Kiran Kumar.(2021). “Early estimation model for 3D-discrete indian sign language recognition using graph matching.” *Journal of King Saud University – Computer and Information Sciences 33*, 852–864. https://doi.org/10.1016/j.jksuci.2018.06.008

[26] Muhammad Aminur Rahaman, Kabiratun Ummi Oyshe, Prothoma Khan Chowdhury, Tanoy Debnath, Anichur Rahman, Md. Saikat Islam Khan. (2024). “Computer vision-based six layered ConvNeural network to recognize sign language for both numeral and alphabet signs.” *Biomimetic Intelligence and Robotics 4 100141.* https://doi.org/10.1016/j.birob.2023.100141

[27] Nehal F. Attia, Mohamed T. Faheem Said Ahmed, Mahmoud A.M. Alshewimy. (2023). “Efficient deep learning models based on tension techniques for sign language recognition.” *Intelligent Systems with Applications 20 200284.* https://doi.org/10.1016/j.iswa.2023.200284

[28] K. Sharma, K. A. Aaryan, U. Dhangar, R. Sharma, and S. Taneja. (2022). “Automated Indian Sign Language Recognition System Using LSTM models.” *ICCCIS 2022*, pp. 461–466. doi: 10.1109/ICCCIS56430.2022.10037711.

[29] Likhar, P., & Rathna, N. G. (2021). “Indian Sign Language Translation using Deep Learning.” *IEEE Region 10 Humanitarian Technology Conference, R10-HTC, 2021-Septembe*r. https://doi.org/10.1109/R10-HTC53172.2021.9641599

[30] Gautham Jayadeep, Vishnupriya N V, Vyshnavi Venugopal, Vishnu S, Geetha M. (2020). “Mudra: Convolutional Neural Network based Indian Sign Language Translator for Banks.” *ICICCS 2020. IEEE Xplore Part Number:CFP20K74-ART;* ISBN: 978-1-7281-4876-2

[31] Obi, Y., Claudio, K. S., Budiman, V. M., Achmad, S., & Kurniawan, A. (2022). “Sign language recognition system for communicating to people with disabilities.” *Procedia Computer Science,* 216, 13–20. https://doi.org/10.1016/j.procs.2022.12.106

[32] A. Sridhar, R. G. Ganesan, P. Kumar, and M. Khapra. (2020). “INCLUDE: A Large Scale Dataset for Indian Sign Language Recognition.” MM 2020 - *Proceedings of the 28th ACM International Conference on Multimedia, Association for Computing Machinery, Inc, Oct. 2020*, pp. 1366–1375. doi: 10.1145/3394171.3413528.

[33] J. Bora, S. Dehingia, A. Boruah, A. A. Chetia, and D. Gogoi. (2023). “Real-time Assamese Sign Language Recognition using MediaPipe and Deep Learning,” *Procedia Comput Sci, vol. 218,* pp. 1384–1393, 2023, doi: 10.1016/j.procs.2023.01.117.

[34] M. Bansal and S. Gupta. (2021). “Detection and Recognition of Hand Gestures for Indian Sign Language Recognition System.” *IEEE International Conference on Signal Processing,Computing and Control,* pp. 136–140. doi: 10.1109/ISPCC53510.2021.9609448.

[35] Ingoley, S. & Bakal, J. (2023). “Indian Sign Language Recognition Using Hand-Pose Key Points and Transfer Learning.” *International Journal of Applied Engineering & Technology*, Vol. 5 No.4, December, 2023, ISSN: 2633-4828.

[36]Shreya Daga, Atharva Dusane, Dyuti Bobby. (2024). “With You - Indian Sign Language Detection and Alert System”, *ESCI,* 979-8-3503-0661-3/24, DOI: 10.1109/ESCI59607.2024.10497366

[37] Shankara Narayanan V, Sneha Varsha M, Padmavathi S. (2024). “Continuous Sign Language Recognition using Convolutional Neural Network”, *ICETITE*, 979-8-3503-2820-2/24, DOI: 10.1109/IC-ETITE58242.2024.10493715

[38] Sreemathy, R., Turuk, M. P., Chaudhary, S., Lavate, K., Ushire, A., & Khurana, S. (2023). “Continuous word level sign language recognition using an expert system based on machine learning”. *International Journal of Cognitive Computing in Engineering,* 4, 170–178. https://doi.org/10.1016/j.ijcce.2023.04.002

[39] Shahan Ahmed , Sumit Kumar Kar , Sarnali Basak. (2023).”A Novel Approach for Recognizing Real-Time American Sign Language (ASL) Using the Hand Landmark Distance and Machine Learning Algorithms”, *IEEE 9th WIECON-ECE* | 979-8-3503-1965-1