Inter-Sign Language Conversion Using Convolutional Neural Networks (CNNs)

Adrian Calvin K.A., Atchaiyaraj S., Ahmed Sherif Mohammed Rafi, Dr. Maharajan M.S.

Department of Artificial Intelligence & Data Science, Panimalar Engineering College, Bangalore trunk Road,
Nazarethpettai, Varadharajapuram, Chennai-600123.

Abstract:

Sign language is a quintessential communication medium among individuals who are either deaf or hard of hearing. It involves gestures and visual-spatial expressions to convey and communicate with one another. It is the only means of communication for people deprived of oral skills. There is no universally followed standard sign language for communication. In our research, we propose a method that could be used to resolve the issue by converting one sign language to the other.

Our method would help in not only converting sign language to text but also to another sign language that would help people who are deprived of text knowledge. We use a novel Convolutional Neural Networks (CNNs) architecture to recognize the sign language and display the gesture of the target sign language.

Keywords: Convolutional Neural Networks (CNNs), Source sign language, Target sign language, Visual-spatial expression.

Related Works:

People with hearing and vocal disabilities often struggle to communicate using technology, this paper [1] explores existing research and advancements in the field of assisting with communication. It focuses on a novel system designed to enhance communication for this population by combining sign language recognition and speech-to-sign conversion techniques. This paper [1] demonstrates the system's methodology as an innovative approach to sign language recognition, which effectively identifies and translates signs from video sequences characterized by minimal clutter and dynamic backgrounds. This is achieved through the implementation of skin colour segmentation techniques. The system's ability to differentiate between static and dynamic gestures and extract appropriate feature vectors is emphasized. [1]

Using deep learning this paper [2] to the understanding of object detection advancements by categorizing algorithms into two-stage and one-stage detectors, In the category of two-stage detectors, several algorithms have been instrumental in bolstering detection accuracy. These include the Region-based Convolutional Neural Network (RCNN), its successors Fast RCNN, and Faster RCNN. These methods emphasize precision and accuracy in object detection. [2]

This paper [3] works on the advances of ASL identification by suggesting an innovative Convolutional Neural Networks (CNNs)-based model that improves accuracy. This study intends to address a vital part of communication for the hearing impaired by enhancing and improving existing techniques. The results described in this work show how modern machine learning approaches have the potential to significantly increase the precision of ASL gesture detection, enabling more efficient and meaningful interaction between members of the deaf and hearing populations. [3]

By the usage of a Convolutional Neural Networks (CNNs)-based system for Arab Sign Language identification, this research study contributes to the area of automatic sign language recognition. The paper [4] proves the efficacy of the suggested strategy by using knowledge from earlier research on recognition systems, deep learning, and conventional machine learning methods. The comparative study's findings support the system's potential for social and humanitarian effects, highlighting its importance in fostering inclusion and accessibility for the deaf-dumb population. [4]

The process of turning sign language into text or voice to improve communication between deaf-mute people and the general population. Due to the complex and varied hand movements used in sign language, this activity is of great societal significance yet is nonetheless difficult. By putting forth a novel Convolutional Neural Networks (CNNs)-based model that automatically extracts spatial-temporal information from unprocessed video streams, this study contributes to the area of sign language recognition. [5]

The field of sign language recognition systems and its use in aiding the community of people with speech and hearing impairments. Accurately recognizing sign motions is a significant difficulty that is being addressed by this project. Previous studies have shown that it is possible to teach computers how to read signs; however, the accuracy of this training depends on how well the categorization and prediction procedures aided by machine learning techniques perform. By building on the groundwork established by earlier research on Convolutional Neural Networks (CNNs) parameter tweaking, gesture detection, and machine learning a deeper knowledge of Convolutional Neural Networks (CNNs) architecture optimization results from the analysis of optimal filter sizes. [6]

The challenges faced by people with hearing loss and the crucial role sign language is playing in removing these barriers. With the rise of sign language, people with hearing and speech disabilities now have a powerful form of communication that makes it easier for them to communicate with others and integrate into society. The work expands on earlier investigations into deep learning, computer vision, and gesture detection as Convolutional Neural Networks (CNNs), and Computer Vision. The outcomes highlight the viability and efficacy of the suggested method in precisely identifying sign motions, hence enabling more effective communication for those with hearing difficulties. [7]

The visual interpretation of sign language presents a variety of challenges in the field of computer vision due to the distinctive signs' subtle variations in hand form, motion profile, and spatial arrangements of the hand, face, and body parts. This paper [8] presents a thorough analysis of the dynamic environment of deep learning-based sign language recognition like Convolutional Neural Networks (CNNs). The survey contextualizes the evolution of this study field by citing fundamental works, recent developments, and problems that still need to be solved. [8]

This paper explains the recommended method within the context of hand gesture detection and explores the implications for automated systems using the Convolutional Neural Networks (CNNs). By referencing past studies on gesture recognition, image enhancement, segmentation techniques, and Convolutional Neural Networks (CNNs) classification, the study creates a strong foundation for the suggested strategy. The experimental outcomes validate the methodology's efficacy and are in line with advancements made in the broader field of image classification and computer vision. [9]

Due to the general public's prevailing ignorance of sign language, academics have looked at technology methods to improve communication with the deaf community. Technology offers a promising way to function as a bridge and enable successful communication and interactions as it develops further. The technique is positioned within the larger context of communication options for people with hearing impairments in the research paper's [10] section on similar publications. The work builds a thorough basis for the proposed technique by drawing on earlier research on technology as a communication bridge, image processing, hand key point libraries, Convolutional Neural Networks (CNNs), and algorithmic fusion. The experimental findings support the Ensemble method's applicability and efficiency, and they are consistent with emerging trends in algorithmic fusion and improving communication for the deaf community. [10]
A table that compares the advantage of using Convolutional Nueral Networks is shown below:

Advantages	CNNs vs Traditional Techniques
Image Classification Accuracy	Significantly higher accuracy in tasks like image classification and recognition.
Feature Learning	Automatically learn hierarchical features from raw image data, reducing the need for manual feature engineering.
Parameter Efficiency	Share parameters through convolutional kernels, leading to more parameter-efficient models
Deep Architectures	Improved performance with increasing model depth, capturing more complex image representations.
Robustness to Noise	Handle noisy or cluttered images better due to learned meaningful features.
State-of-the-Art Results	Consistently outperform other techniques in competitions and challenges.
Transfer Learning	Support transfer learning, allowing pre-trained models to be adapted to new tasks with small datasets.
Scale Invariance	Effective at handling variations in object scale through multi-scale feature extraction.
Semantic Segmentation	Excel in semantic segmentation tasks, providing pixel-level object delineation.
Real-Time Processing	Enable real-time image processing for applications like object detection and autonomous vehicles.

Table:1, Advantages of using CNNs over traditional techniques.

Introduction:

Sign language is a visual language that uses hand movements, facial expressions, body language, and other visual cues to communicate meaning. To interact with the hearing population and each other, it is largely utilised by deaf or hard-of-hearing people. Sign languages are fully formed languages with their syntax and vocabulary, much like spoken languages.
The practice of translating one sign language into another is known as "sign language interpretation" or "sign language translation." Regional differences exist in sign languages

just as we do in spoken languages. While the fundamentals of sign language are universal, there can be substantial regional variation in the signals and expressions utilised.

The interpreter or translator needs to be proficient in both the source sign language (the language being signed by the speaker) and the target sign language (the language to which it needs to be converted), which would require a deep understanding of the grammar and vocabulary of both the sign languages.

Word-for-word translation from sign language is not always possible. Instead, the interpreter concentrates on communicating the message's main ideas and meaning while modifying the signs, gestures, and facial expressions to meet the customs of the target sign language.

To overcome the above-mentioned problem, we have implemented our research in the field of machine learning and have come up with a solution, to resolve it. In our research, we focus on implementing a fully automated translator which would be handy most of the time, rather than a human interpreter or a translator.

To automatically translate or interpret sign language movements from one language to another, machine learning is used to convert one sign language to another. Due to the visual character of sign languages and the differences in gestures, emotions, and body movements between different sign languages, the procedure we have undertaken, is challenging.

Proposed System:

Architecture Diagram:

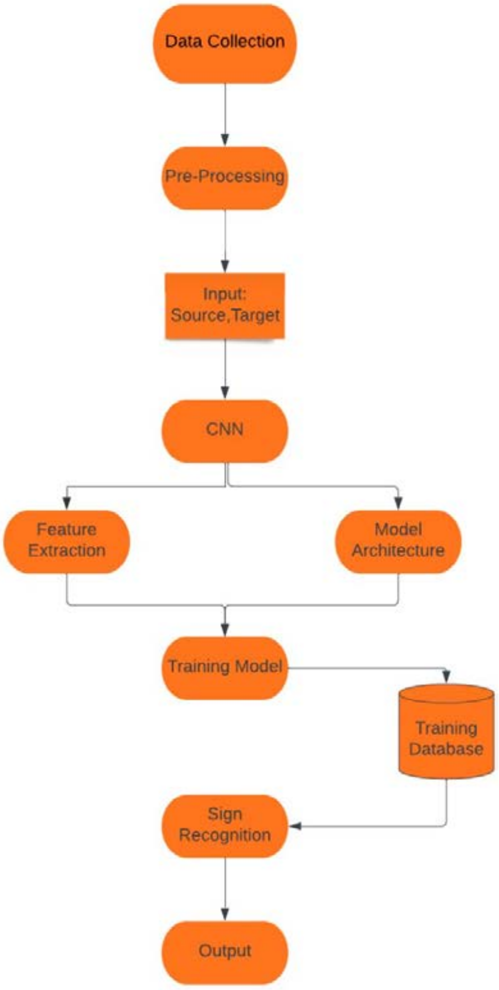

fig:1, Overall architecture Diagram

The algorithm for the entire processes that are involved are given as follows:

Step-1: Gather and label a dataset of the sign languages used.
Step-2: Use the labeled dataset to train a CNN model.
Step-3: Provide an image of the sign you wish to transform as the input image.
Step-4: Resize and normalize the input image.
Step-5: Predict the sign category using the trained CNN.
Step-6: Interpret the prediction and identify the target sign in the post-processing stage.
Step-7: Display or store the converted sign image as the output image.

Data collection and pre-processing:

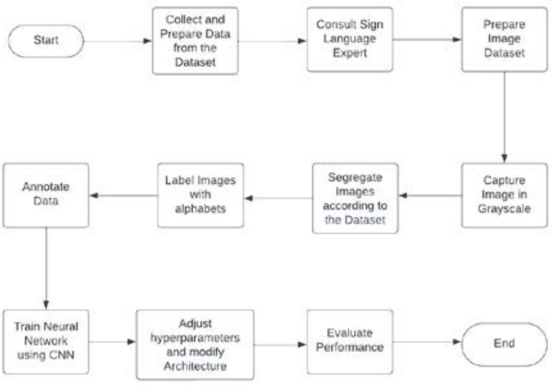
Start
Collect and Prepare Data from the Dataset
Consult Sign Language Expert
Prepare Image Dataset
Capture Image in Grayscale
Segregate Images according to the Dataset
Label Images with alphabets
Annotate Data
Train Neural Network using CNN
Adjust hyperparameters and modify Architecture
Evaluate Performance
End

fig:2, Steps involved in Data Pre-processing

To train and fit any model, firstly data must be collected and made processing-ready, as in fig. [2], to proceed further into the development of the model. A sign language expert is consulted to prepare the data required for training, testing and validation of the model. The data collected is in image format. A dataset consisting of images of an individual letter of the sign language is prepared. The images are captured under different lighting and different angles to increase the accuracy of the model.

Images that are collected are all formatted in grayscale to increase computational efficiency as it has only one single channel but the regular RGB images have three channels. Grayscale images are also less prone to noise in the channels and have fewer dimensions that reduce the complexity while processing. Grayscale is best suited for texture and shape analysis.

All the images are then segregated according to the meaning of the gesture, in other words, images are placed in directories for the corresponding English alphabet they mean. Individual images in the directories have proceeded next for labelling. Labels are created in English alphabets. Images are labelled because Neural Networks are trained using supervised learning in which the model learns to map the data to the corresponding output label. Labelling also helps in fine-tuning the model and analysing how decisions are made using a particular model.

After labelling the data is annotated with each dataset to the corresponding labels for each sign gesture. Annotations serve as references or guides for the model for adjusting their internal parameters. They are essential to adjusting hyperparameters, modifying the model's architecture and improving its overall performance.

The steps that are undertaken Data collection and pre-processing are given below:

Step-1: Obtain a collection of images that include the signs you wish to convert.

Step-2: Label and annotate the images with the appropriate sign categories.

Step-3: Create training, validation, and test sets from the dataset.

Step-4: To make your training dataset more diverse, perform data augmentation. Transformations like rotation, scaling, and flipping are frequently used.

Training the model:

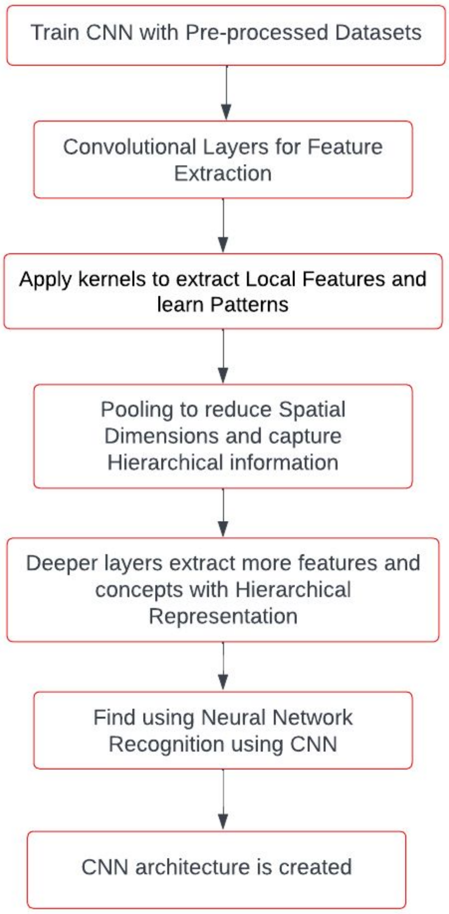
Train CNN with Pre-processed Datasets
Convolutional Layers for Feature Extraction
Apply kernels to extract Local Features and learn Patterns
Pooling to reduce Spatial Dimensions and capture Hierarchical information
Deeper layers extract more features and concepts with Hierarchical Representation
Find using Neural Network Recognition using CNN
CNN architecture is created

fig:3, Steps involved in Training the model

The required datasets are pre-processed and now it’s ready to be fed as input for training the model, as in fig. [3]. Since we need to process images or frames, Convolutional Neural

Networks (CNNs) architecture is used. The advantage of using Convolutional Neural Networks (CNNs) is that they are specifically designed to analyse visual data.

Convolutional Neural Networks (CNNs) use convolutional layers to extract local features and to learn automatically from the input data and hence by helping to capture patterns like corners, textures and the remaining elements present in an image. Convolutional Neural Networks (CNNs) play a vital role in understanding the contents of the image.

Each convolutional layer has a set of filters or kernels that are applied to the image. To produce feature maps the filter or kernel slices across the image while executing pixel-wise multiplication and summation through which each pattern or characteristic feature of a particular image is recognized.

Pooling is also a technique that is used in Convolutional Neural Networks (CNNs) to retain the important contents of the image whilst reducing the spatial dimensions of the feature maps. Lowering the spatial dimensions and pooling aids the networks in capturing the hierarchical information. It is very helpful for identifying patterns at various scales.

As multiple layers make up the Convolutional Neural Networks (CNNs) the deeper one moves into the network the layers progressively learn more abstract and sophisticated information about the image. The networks can comprehend linkages and concepts present in the data at much higher levels because of hierarchical representation.

Compared to fully connected networks, the number of learnable parameters is lower in convolutional layers, since shared weights are used. This makes the model more efficient while handling large datasets. In a variety of image-related tasks, such as image classification, and object detection Convolutional Neural Networks (CNNs) have attained cutting-edge performance. For our choice of architecture and dataset Convolutional Neural Networks (CNNs) based solution would play a critical role in the success of our model.

A set of mathematical procedures can be used to express the fundamental elements of a CNN. The fundamental components of a CNN are represented by the following abbreviated formula:

$$\mathbf{O=ReLU((i*K)+b)}$$

In the above shown formula,

- The input image is represented by 'i'.
- The learnable convolutional kernel (filter) is represented by 'K'.
- '*' represents the "convolution operation" used.
- 'b' stands for the bias term.
- ReLU(x) is the Rectified Linear Unit activation function.

The above-mentioned processes are shortly briefed as follows:

Step-1: The CNN architecture that is appropriate for the task is created. Convolutional layers, pooling layers, fully connected layers, and an output layer are frequently present.

Step-2: A suitable loss function is picked for the classification task, such as categorical cross-entropy.

Step-3: The training dataset is utilized to train the CNN model.

Step-4: To keep an eye on training progress and avoid overfitting, the validation dataset is used.

Step-5: The model is trained repeatedly until it reaches an acceptable level of accuracy.

Step-6: The test dataset are used to assess the trained model's performance on new data.

Step-7: Measurements including accuracy, precision, recall, and F1-score are computed.

Step-8: The trained model is saved to a file after a commendable performance could be extracted form the model, for future use.

Output generation:

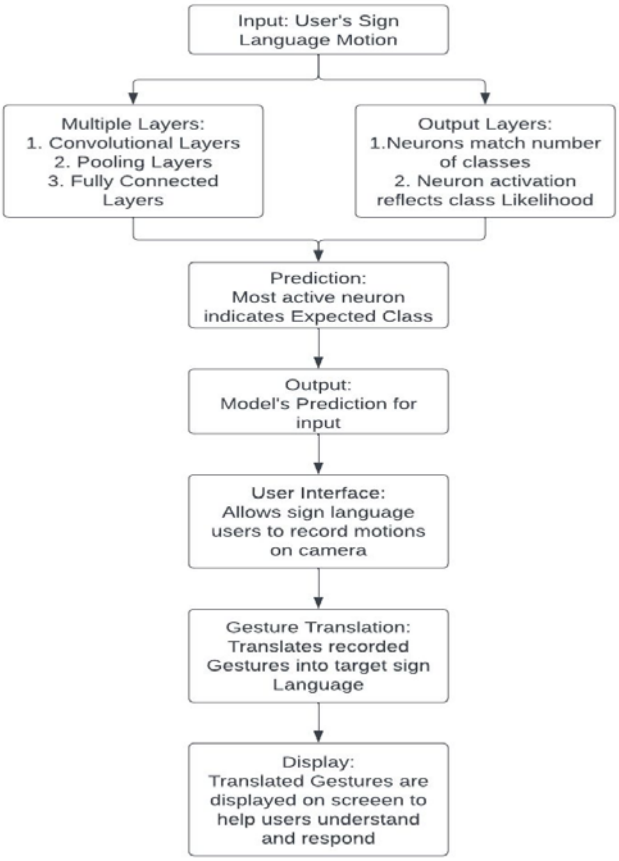
Input: User's Sign Language Motion
Multiple Layers: 1. Convolutional Layers 2. Pooling Layers 3. Fully Connected Layers
Output Layers: 1.Neurons match number of classes 2. Neuron activation reflects class Likelihood
Prediction: Most active neuron indicates Expected Class
Output: Model's Prediction for input
User Interface: Allows sign language users to record motions on camera
Gesture Translation: Translates recorded Gestures into target sign Language
Display: Translated Gestures are displayed on screeen to help users understand and respond

fig:4, Procedure for output generation

Among the several layers present in Convolutional Neural Networks (CNNs) the output layer is also present at the end. It is finally reached from the completely linked layers. The number of classes in the classification issue is matched by the neurons present in the output layer. The degree or measure to which a certain neuron is activated reflects the likelihood that the input falls into a certain class.

The process determines how the output layer's activation function operates. This softmax() is frequently used to convert raw scores into probability distributions over classes in a multi-class classification.

Predictions are involved in the output layer neuron. The neuron that is most active in the output layer indicates the expected class. The model's prediction for the input is the class label to which the particular neuron belongs.

A simple user-friendly interface is developed to enable sign language users to record their motions on camera and feed them. To help the users comprehend and react appropriately the gestures are translated into target sign language and displayed on the screen as output, as in fig. [4].

Experimental Results And Discussion:

Pre-processing the data:

Pre-processing is a crucial step in preparing image datasets for sign language conversion. It is the most essential step that is followed while training Convolutional Neural Networks (CNNs). Good pre-processing always serves to improve the quality of data and the training process. A great amount of model performance is achieved through proper pre-processing.

Firstly we create a directory called dataset inside which there are two more sub-directories namely "source directory" and "target directory". In our case, we choose the source directory to be our source language dataset or the sign language in which the input is to be received and the same is the case for the target directory. Both the source and target directory contain many folders with each folder representing an alphabet. Inside each folder, there are 3,000 images of the sign gesture the folder represents.

With accuracy as the main metric, we evaluated how well various model architectures performed in image processing tasks. The best appropriate model architecture is determined by a variety of factors, including the availability of computational resources and the trade-off between accuracy and efficiency. MobileNet displayed efficiency while ResNet attained the maximum level of accuracy, making it a good contender for embedded and mobile applications. The particular requirements and limitations of the task at hand should guide the choice of the optimal model. The above comparison is visually briefed in the fig.[5].

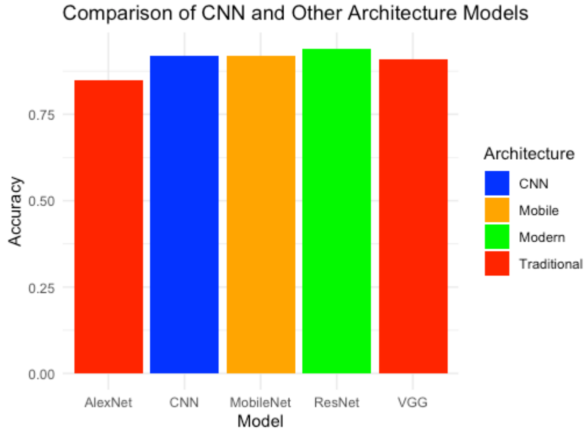

fig:5 Accuracy Comparison Chart

After all the images are loaded in their respective directories/folders they all are to be resized as Convolutional Neural Networks (CNNs) architecture is compatible with only 224x224 pixels or 256x256 pixels. This process would ensure that all the images are compatible with the network's architecture.

Once resizing is done the dataset must be augmented using transformations such as scaling, flipping, rotating and translating. It makes it easier for the model to generalise to new images and also increases the diversity of the training data.

The datasets are now ready to be split into training, testing and validation. For each purpose, the datasets are split in a ratio of 8:1:1 respectively (i.e.) out of 100 images 80 images are used for training, 10 images are used for testing and 10 images are used for validation. Finally, the pre-processed datasets are saved in a suitable format that would help load the data while training the model.

Test setup:

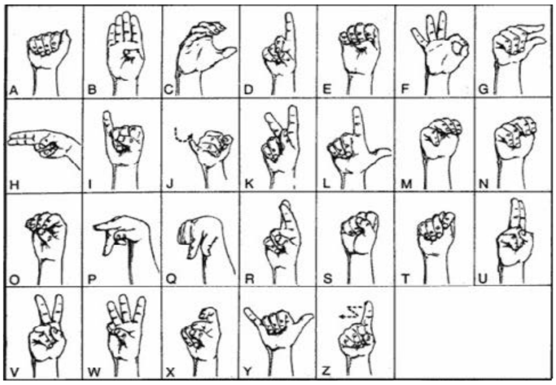

fig:6 Source sign language gestures

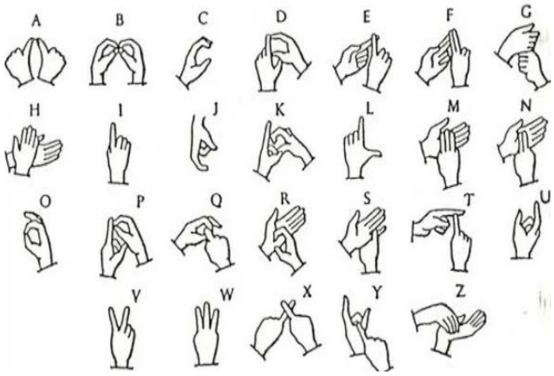

fig:7 Target sign language gestures

For testing our development, a laptop with MAC OS with an M1 processor and 4 GB of RAM is used. The datasets for 5 sign languages are pre-processed and saved beforehand. An expert in the field of sign language is consulted for advice and help. The expert chooses the source(fig:6) and target(fig:7) sign language of their choice.

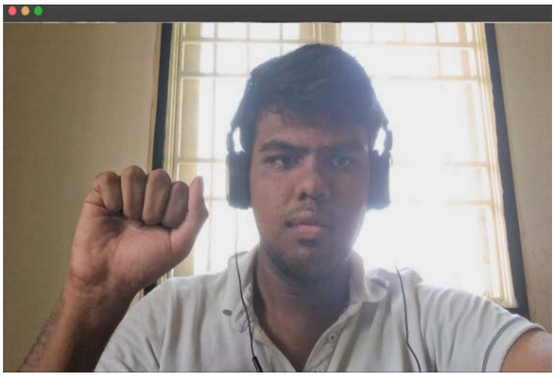

fig:8 Input gesture

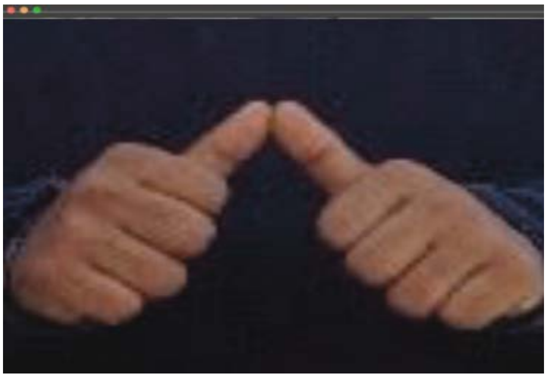

fig:9 Output

After the source and target languages are received, the web camera present on the laptop becomes functional as the camera light turns 'ON' simultaneously along with the tab that displays the camera input. The expert makes a gesture representing a letter from the source language, as in fig. [8], The model identifies the gesture and the camera tab is closed. A new tab is opened, as in fig. [9], that consists of a sign language gesture as output. The expert verifies the output and confirms the output to be legitimate.

Conclusion:

A possible method for overcoming communication hurdles and enabling effective translation across various sign languages is using Convolutional Neural Networks (CNNs). Convolutional Neural Networks (CNNs) provide a strong foundation for identifying and converting sign language motions because of their natural capacity to learn hierarchical patterns from visual data.

The deaf and hard-of-hearing people may benefit from cross-language communication thanks to the use of Convolutional Neural Networks (CNNs) for sign language conversion. It can improve diversity, facilitate accessibility, and remove linguistic obstacles.

In conclusion, the use of Convolutional Neural Networks (CNNs) for sign language conversion has the potential to promote communication equality and deaf people's sense of empowerment. However, developing practical, accurate, and respectful solutions that genuinely improve interlanguage communication requires a careful, team-based, and culturally responsible approach.

References:

[1] Kumar, Anup, Mevin. “Sign Language Recognition.” *2016 3rd International Conference on Recent Advances in Information Technology (RAIT)*, IEEE, Mar. 2016.

[2] Adarsh, Pranav, Pratibha Rathi, et al. "YOLO v3-Tiny: Object Detection and Recognition using one stage improved model." In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pp. 687-694. IEEE, 2020.

[3] M. M. Rahman, M. S. Islam, M. H. Rahman, et.at., "A New Benchmark on American Sign Language Recognition using Convolutional Neural Network," *2019 International Conference on Sustainable Technologies for Industry 4.0 (STI)*, Dhaka, Bangladesh, pp. 1-6, 2019.

[4] S. Hayani, M. Benaddy, O. El Meslouhi, et al., "Arab Sign Language Recognition with Convolutional Neural Networks," *2019 International Conference of Computer Science and Renewable Energies (ICCSRE)*, Agadir, Morocco, pp. 1-4, 2019.

[5] Uyyala, Prabhakara. "SIGN LANGUAGE RECOGNITION USING CONVOLUTIONAL NEURAL NETWORKS." *Journal of Interdisciplinary Cycle Research* 14.1, pp. 1198-1207. 2022.

[6] Jain, V., Jain, A., Chauhan, A. *et al.* “American Sign Language Recognition Using Support Vector Machine and Convolutional Neural Network.” *International Journal of Information Technology*, vol. 13, no. 3, Springer Science and Business Media LLC, pp. 1193–200. 2021.

[7] Murali, Romala Sri Lakshmi, L. D. Ramayya, and V. Anil Santosh. "Sign language recognition system using convolutional neural network and computer vision." *International Journal of Engineering Innovations in Advanced Technology*, vol. 4, no. 4 (2020).

[8] Rastgoo, Razieh, Escalera. “Sign Language Recognition: A Deep Survey.” *Expert Systems With Applications*, vol. 164, Elsevier BV, p. 113794, 2021.

[9] Neethu, P. S., et al. “An Efficient Method for Human Hand Gesture Detection and Recognition Using Deep Learning Convolutional Neural Networks.” *Soft Computing*, vol. 24, no. 20, Springer Science and Business Media LLC, pp. 15239–48, 2020.

[10] C. Suardi, A. N. Handayani, R. A. Asmara, et al., "Design of Sign Language Recognition Using E-CNN," *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, Surabaya, Indonesia, pp. 166-170, 2021.

[11] Reagan L. Galvez, Argel A. Bandala., et al., “Object Detection Using Convolutional Neural Networks”, IEEE, 2018.

[12] Md. Mehedi Hasan Naim, Rohani Amrin, et al., “Object Detection from Images using Convolutional Neural Network based on Deep Learning.” Volume 09, Issue 09, September 2020.

[13] Abhinav Juneja, Sapna Juneja, et al., “Real Time Object Detection using CNN based Single Shot Detector Model.” *Journal of Information Technology Management*, Vol.13, No.1, 2021.

[14] Tausif Diwan, G. Anirudh & Jitendra V. “Object detection using YOLO: challenges, architectural successors, datasets and applications.”, *Tembhurne Multimedia Tools and Applications*, volume 82, pages 9243 – 9275, 2023.

[15] Aishwarya Sarkale, Kaiwant Shah, et al., “A Literature Survey: Neural Networks for object detection”, *VIVA-Tech International Journal for Research and Innovation*, Volume 1, Issue 1, 2018.

[16] Ashwani Kumar, Sonam Srivastava. “Object Detection System Based on Convolution Neural Networks Using Single Shot Multi-Box Detector”, *Procedia Computer Science*, Volume 171, Pages 2610-2617, 2020.

[17] Junsong Ren, Yi Wang. “Overview of Object Detection Algorithms Using Convolutional Neural Networks”, Vol.10 No.1, January 2022.

[18] Jun Deng, Xiaojing Xuan, et al., “A review of research on object detection based on deep learning”, *Journal of Physics: Conference Series*, *Ser.* 1684 012028, 2020.

[19] Sharvani Srivastava, Amisha Gangwar, et al., “Sign Language Recognition System using TensorFlow Object Detection API”, *Computer Vision and Pattern Recognition*, volume 2, 2022.

[20] Reddygari Sandhya Rani, R Rumana, et al., “A Review Paper on Sign Language Recognition for The Deaf and Dumb”, Volume 10, Issue 10, October, 2021.