Indian Sign language recognition using Custom CNN

Annapurna Shankar Balikai
Department of ECE
KLE Technological University
Hubli, Karnataka, India
annapurnasb45@gmail.com

Aishwarya Suresh Naik
Department of ECE
KLE Technological University
Hubli, Karnataka, India
ashnaik2003@gmail.com

Tejaswini Shankar Sattikar
Department of ECE
KLE Technological University
Hubli, Karnataka, India
sattikartejaswini@gmail.com

Satish Chikkamath
Department of ECE
KLE Technological University
Hubli, Karnataka , India
chikkamath@kletech.ac.in

Suneeta V. Budihal
Department of ECE
KLE Technological University
Hubli, Karnataka , India
sunneta_vb@kletech.ac.in

***Abstract*—Communication is very crucial in day-to day life mainly when it comes to education, employment, exchanging of ideas and feelings. Normally humans use speech as their mode of communication but for differently abled people sign language is the mode of communication as they cannot speak or listen, which creates the communi cation barrier. The proposed model focuses on converting the Indian sign language into multilingual text and audio using Mobilenetv2 and translation API. For the dataset videos are recorded and curated by keeping the reference of Indian Sign Language (ISL). The datasets consist of videos of 20 English words. All videos have either white or black background with duration of 10 seconds and the resolution of videos is 1080 pixels and 30 frames per second. Each word contains 5 videos demonstrated by 5 different people. The total number of frames of entire dataset is 30,000. The model achieved the accuracy of 90**

***Index Terms*—Sign language, Indian Sign language, con volutional neural networks (CNN),long short-term mem ory (LSTM), Mobilenet.**

I. Introduction

Sign language is the fundamental mode of communication for differently abled people. According to resources, there are around 1.4 billion people in India. Among them, around 50 million people are differently abled who use sign language as their mode of communication, which creates a communication barrier between the differently abled people and the people who do not understand sign language. Communication barriers lead to challenges like limited access to education, cultural disconnect, and dependence on intermediaries, which leads to a lack of access to public resources and job opportunities. To overcome the communication barrier, there is a need for the interpretation of sign language, so many research works have been done on sign language to text conversion. Existing research works are focused on predicting alphabets (A-Z), numbers (0-9), and some have been done word-level, but the prediction of words is limited to only one language. Since India is a multilingual country, different sign languages are used in different regions, which makes standardization of translation of sign language difficult.According to literature survey all the existing papers have only converted the sign language into text which becomes difficult for illiterate people to read.

Sign language recognition (SLR) models are built for the interpretation of sign language into text using deep learning algorithms like convolutional neural networks (CNN) and YOLO. Existing models take images of signs as input, which mainly focus on static images of hand gestures. The proposed model takes videos of signs as input and converts them to text using a custom CNN algorithm. The text is given to any translation API for converting the text into the required language, which helps to standardise the translation of sign language. and also the text is converted into audio of respective languages

II. Literature Review

Existing work [1] has been done for the recognition of Indian sign language (ISL) using neural networks, Single Shot Multibox Detector (SSD) MobileNet is an efficient object detection algorithm that identifies the object in a single step. MobileNet is used to extract the features from input images; the algorithm is ideal for real-time application in the aspect of speed and accuracy. The dataset comprises 26 classes (A-Z letters from ISL) and is given as an input to the pretrained model of SSD-MobileNet, which is fine-tuned and achieved an accuracy of 96.11 mean Average Precision (mAP) at 50IoU and 82on MobileNet and Inception V3 for real-time recognition of Indian Sign Language (ISL). They developed a dataset comprising 2,820 images across 47 classes, which included

26 alphabets, 10 numerals, and 11 gestures. A comparison between the MobileNet and InceptionV3 models revealed that MobileNet achieved higher accuracy, with a remarkable 99accuracy for both training and testing phases, and they have built a mobile application called SIGNify recognising static sign language gestures.

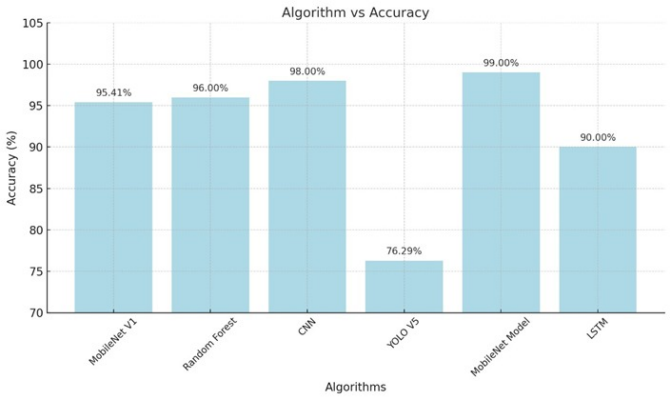

Fig. 1: Algorithm vs Accuracy

Another noteworthy work [5] focused on both static images and short video clips of American Sign Language (ASL). The system employed YOLO (You Only Look Once) for recognising static images and Long Short-Term Memory (LSTM) networks for dynamic video clips. YOLO demon strated excellent performance in detecting and recognising individual signs from static images. For dynamic gestures, the keypoints are extracted using Mediapipe. LSTM effectively processed sequential data to achieve an accuracy of 92.3for dynamic gestures. In this context [7], MobileNetv1, a lightweight convolutional neural network, uses the depth-wise convolution to reduce the computations. Input images sized to [64*64] pixels were applied to the pretrained MobileNet model on ImageNet, and the last layer was replaced and fine tuned to classify 29 sign language gestures. The dataset lacks information on variations in background, which are critical for real-world applications.

The study [8] takes input from a live webcam using Me diapipe, which is a critical tool that captures the keypoint from the dataset. The random forest algorithm is used to classify gestures, and classified gestures are shown as text and converted into speech using a tool called GTTS (Google Text to-Speech). With an average of processing 0.3 seconds per gesture, the model works well but struggles with faster or more complex gestures. The data [9] from the live webcam is pre processed using a Gaussian filter, which reduces noise, and features from the processed hand image are extracted using image analysis methods. The extracted features are applied to a deep learning neural network called a convolutional neural network (CNN). After the gesture recognition, it is mapped to its corresponding character. These characters can be combined to form words, and these texts are converted to speech using Google Text-to-Speech (GTTS) API. The accuracy of the algorithm is as high as 99.85%, allowing the system to operate efficiently in real-time scenarios. A similar research study [3] for bridging the communication gap was done using you only look once (YOLOV5), taking the 35 common videos as the input to it, where the videos are preprocessed into frames that were labelled, and each video is having different backgrounds and signers to improve the diversity and achieve the training accuracy of 76.3% and the testing accuracy of 51.44%. The low video quality or poorly positioned signs (e.g., too close or too far from the camera) reduced accuracy. The data collection in study [8] and [9] is done in the same way in study [6] through webcam. Using Mediapipe, key features (hand and body position) are extracted. The static gestures are fed as input to a convolutional neural network model (CNN), and dynamic gestures are fed to long short-term memory (LSTM), and the detected signs are combined into sentences using a language model like Google Flan T5. The accuracy achieved for static gestures is 99.2%, and for dynamic gestures, it is 90.08%. The study [11] worked on the Word-Level American Sign Language dataset, which has the video clips with labels for each sign, and input data is processed using the media pipe and given to the CNN and RNN algorithms for keeping the context of the temporal information of dynamic signs and NLP from the TensorFlow library to form sentences.

III. Methodology

The main objective of model is the prediction of text for given video of particular word sign and then the predicted text is given to any translation API for translation into mul tiple languages and further text is converted to audio of the particular language.

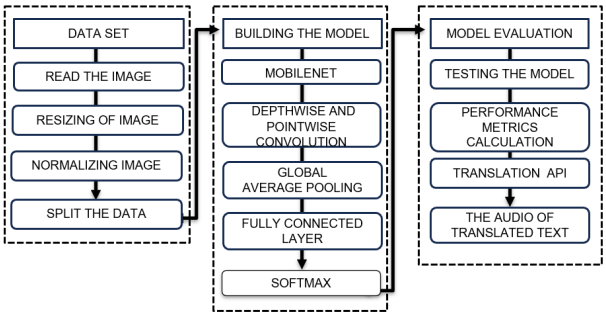

Fig. 2: The proposed model

A. A. Dataset

According to literature survey the existing datasets of sign language are specific to particular regional signs which are not standard and some datasets available in official website of Indian sign language the videos have different backgrounds and low resolution which creates difficulty for model to getA patterns. To address these challenges dataset is curated by referring the official website of Indian sign language (ISL) [13]. The datasets consist of videos of 20 English words. The videos are recorded by referring to the official website of

Indian Sign language. All videos have either white or black background with duration of 10 seconds and the resolution of videos is 1080 pixels and 30 frames per second. Each word contains 5 videos demonstrated by 5 different people. The total number of frames of entire dataset is 30,000.

Fig. 3: Videos of the signs of following English words 1. Accommodation 2. Allergy 3. Competition 4. energy 5. book 6. come

B. Preproccessing

In the preprocessing, videos are resized into fixed time intervals to extract a maximum number of frames that capture the dynamic transitions of each sign. This approach ensures that frames representing the transitions between signs are preserved, allowing the model to learn the full context of the dynamic gestures. The extracted frames are resized to 224x224 pixels to match the input size of MobileNetV2 to ensure compatibility with the model's architecture.

C. Feature extraction

For feature extraction the MobileNetV2 is implemented which is pretrained on ImageNet. MobilenetV2 which is trained on ImageNet identifies a different range of patterns in frames of video. The pretrained weights helps in recognising important patterns in sign language frames, such as edges, textures and shapes allowing the model to focus on extracting higher-level features specific to sign language gestures.

D. Global average pooling

After feature extraction, global average pooling is applied on the extracted features to reduce the dimensionality by averaging the spatial features across the height and width of the feature maps. This process combines the multi-dimensional features into a 1D vector, which is then given to a dense layer. The dense layer of size of 512 to capture higher-level abstract representations of the input data for predictions of sign language.

E. Dropout layer

After global average pooling 50to avoid overfitting, then applying SoftMax to get outputs probabilities for each class.

F. output

Text of particular sign is predicted which is given to translation API and translated text is converted into audio.

IV. RESULTS

The pretrained MobileNetV2 model is loaded and fitted with batch size 16 and 20 epochs. The model detects the sign on which it is trained and the predicted text is translated into multiple languages and converted to audio respectively.

```
Overall video prediction: Energy
Translations for 'Energy':
  hi: ऊर्जा
  kn: ಶಕ್ತಿ
  mr: ऊर्जा
  te: శక్తి
  ta: ஆற்றல்
  ml: ഊർജ്ജം
  bn: শক্তি
  ur: توانائی
  or: ଶକ୍ତି
  as: শক্তি
  gu: ઊર્જા
  bho: ऊर्जा
```

Fig. 4: 1.Hindi,2.kannada,3.marati,4.Telugu,5.Tamil,6. Malyalam,7. Ben gali,8. Urdu, 8. Odia,9. Assamese, 10 . Gujarati, 11. Bhojpuri

The dataset is split into the training dataset of 80% and validation dataset of 20%. The proposed model achieved the training accuracy of 86% and validation accuracy of 90% and validation loss of 32%.

Fig. 5: Train and validation accurcy

Fig. 6: Train and validation loss

REFERENCES

[1] AA Son Kamble, RD Chavhan, BS Jadhao, SM Rathod , "Real-Time Indian Sign Language Detection using SSD Mobilenet", Sardar Patel International Conference on Industry 4.0-Nascent, 2022•ieeexplore.ieee.org year 2022.

[2] X Han, F Lu, G Tian ICETIS 2022, "Sign Lan guage Recognition Based on Lightweight 3D MobileNet-v2 and Knowledge Distillation", 7th International Conference on Electronic Technology, 2022•ieeexplore.ieee.org year 2022.

[3] Tazalli, ZA Aunshu, SS Liya, M Hossain, Z Mehjabeen ,MS Ahmed, MI Hossain 2022, " Computer vision-based Ben gali sign language to text generation", IEEE 5th International Conference on Image Processing ..., 2022•ieeexplore.ieee.org year 2022.

[4] SM Siddiq, S Roopashree, M Suha, MRS Ruthvik, K Divyasree 2021 "SIGNify–A mobile solution for In dian sign language using MobileNet architecture" 2nd Global Conference for Advancement in Technology (GCAT), 2021•ieeexplore.ieee.org.

[5] U Jana, S Paul, D Bhandari 2024 "Real-Time Cap tion Generation for the American Sign Language Using YOLO and LSTM" IEEE International Conference on Infor mation Technology ..., 2024•ieeexplore.ieee.org [6] SX Thong, ELTan, CP Goh 2024 "Sign Language to Text Translation with Computer Vision: Bridging the Communication Gap" 3rd International Conference on Digital Transformation and, 2024•ieeexplore.ieee.org

[7] TN Abu-Jamie, SS Abu-Naser 2022] "Classification of sign-language using MobileNet-deep learning" •philpapers.org

[8] P Jeevanandham, A Hariharan, G Keerthana 2024 "Real Time Hand Sign Language Translation: Text and Speech Conversion" 7th International Conference on Circuit Power and Computing ..., 2024•ieeexplore.ieee.org

[9] A Maitrey, V Tyagi, K Singhal, S Gupta 2023 "A Framework for Sign Language to Speech Conversion Using Hand Gesture Recognition Method" International Conference on Computational Intelligence ..., 2023•ieeexplore.ieee.org

[10] S Kankariya, K Thakre, U Solanki, S Mali, A Chu nawale 2024 "Sign Language Gestures Recognition using CNN and Inception v3" International Conference on Emerging Smart Computing and ..., 2024•ieeexplore.ieee.org

[11] SA Ahmed, YJ Sheriff, B Prakash, SBT Na ganathan 2024 "Sign Language To Text Translator: A Se mantic Approach With Ontological Framework" 10th In ternational Conference on Communication and Signal ..., 2024•ieeexplore.ieee.org

[12] Technical approaches to Chinese sign language pro cessing: A review SM Kamal, Y Chen, S Li, X Shi, J Zheng- IEEE Access, 2019- ieeexplore.ieee.org

[13] https://indiansignlanguage.org/