Literature Survey: Sign Language Translator

Sujay R[1], Somashekar M[2], Aruna Rao B P[3]
[1]*B.E. Student, ECE Dept., K S Institute of Technology*
[2]*B.E. Student, ECE Dept., K S Institute of Technology*
[3]*Assistant Professor, ECE Dept., K S Institute of Technology*

***Abstract*— Communication is a medium for people to exchange feelings and thoughts. The deaf and mute community are withheld from expressing themselves as they cannot talk or speak in regional languages Communication is one of the most essential aspects of human life. Communication is how human beings interact to convey messages, and information and express emotions. Language is the medium through which the process of communication takes place. Language in communication can be verbal, i.e. Using words to read, write and speak or non-verbal i.e., using signs, facial expressions or body language. The deaf and mute community can only communicate using non-verbal means. Sign language is one of the major non-verbal communication methods through which the deaf and mute communicate with others. In order for this community to interact with people who are not aware of Sign language, we propose a system that can convert the Sign language fingerspells to their appropriate words in a standard Language which can be easily understood by others. We can also use the reverse process i.e., to convert written/printed words into their corresponding fingerspells.**

Index Terms: **Sign Language, Human Body Pose, MediaPipe, Machine Learning.**

I.INTRODUCTION

Sign language is a basic means of communication for those with hearing and vocal disabilities. Those disadvantaged face difficulty in their day to day lives to communicate with others. We are aiming to develop a system that would eradicate this barrier in communication with the deaf and mute. Sign language consists of making movements with our hands with certain facial cues. A recognition system would thus have to identify the hand movements, facial expressions and even body pose of the Signer.
American Sign Language is a predominant sign language as it uses a Single hand and most of the fingerspells are static while Indian Sign Language and other languages use two hands and Dynamic Fingerspells. Since the only disability Deaf and Mute people have been communication-related and they cannot use Spoken Languages, hence the only way for them to communicate is through Sign Language. Communication is the process of exchange of thoughts and messages in various ways such as Speech, Signals, Behavior and Visuals, Deaf and mute people make use of their hands to express different gestures to express their ideas to other people. Gestures (Dynamic Fingerspells) are the non-verbally exchanged messages and these gestures are understood with Vision. This non-verbal communication between deaf and mute people is called Sign Language. The Deaf and Mute People face a lot of difficulty in their day to day lives. It becomes, even more, harder for them to travel to places with different native languages. Normally all the Sign Languages are based on the English Alphabet and hence it becomes more challenging for the Deaf and Mute to communicate with the natives as they might not be knowing English
We propose a system that runs on Raspberry Pi and can act as a stand-alone device. We use the Raspberry camera module to capture the fingerspells of the signer and the generated text/sentence is converted to audio with the help of a speaker module that is connected to the 3.5mm audio jack of the Raspberry Pi.
The video captured from the camera module will be processed using OpenCV, and the coordinates of the signer's hand and facial cues are extracted as data points using the MediaPipe library. We will be using these data points to train a deep learning model.

II. PAPER REVIEW

In a previous survey paper, we compared the various techniques used for real-time Sign Language

recognition using Machine Learning. We had compared the Pre-processing techniques of the input video feed, methods used to train the Machine Learning model and the accuracy of each method. The main challenge faced in Sign language recognition is the detection of fingerspells. Various techniques have been employed for hand segmentation, background removal and pre-processingof the images. The other factors affecting the recognition of fingerspells are the colour of the Background, angle of the wrist, and Quality of the camera used.

[1] It is ready to use solution and integrates and works on cross-platformdevices and it is open-sourced. Mediapipe provides face, hand and pose recognition. MediaPipe is a Machine learning model which detects 468 Face landmarks, 33 Full-body landmarks and 21 hand landmarks. These key points define the placement/posture of the person's hands, Body and facial cues in a frame and the coordinates of the key points with respect to the video frame are stored. In order to get better results, we need to provide good quality images with good lighting conditions, brightness and contrast. [7] present a real-time hand tracking solution for generating a skeleton. It recognizes the palm and locates multiple landmarks on the hand. This whole is implemented using the MediaPipe framework. It achieves real time-efficient and increased performance to use in manyreal-time application

[4]3D model of human body pose, hand pose, and facial expression. It captures 2D features and then optimizes model parameters to fit the features during recognition of posture and hand recognition. During the training phase,more numberof cameras are required to capture the pose of the signers as it captures the person in 3d space and optimizes it to 2D (approximately 50 cameras required)

[5]UniPoseis a unified framework for human pose detection and recognition as done by media pipe. It uses a very simple architecture called,"Waterfall" Atrous Spatial Pooling architecture(WASP). It excludes facial expression recognition and only focuses on pose detection

[6] The method used here is dense-flow extraction using a wavelet motion model, facial feature tracking, and edge and line extraction, which captures minute details of the facial expression as it plots more key points on the face,and it excludes the pose detection. The accuracy and efficiency of facial expressionsare more suitable.3D Convolution neural network

In one of the architectures four [5] convolutional layers, five rectified linear units (ReLu), two stochastic pooling layers, one dense and one SoftMax output layer were used and stochastic pooling was considered the best pooling as it showed an increase in accuracy compared to other pooling techniques Sign Language to text and speech using Machine Learning 2020-21

CLACHE (Contrast Limited Adaptive Histogram Equalization) [2] was applied to the dataset images to equalize the lightness in the image frame using the LAB colour system and reduce the noise amplification. Gaussian blurring [2] is used to apply blurring to the image. To obtain the skin, thresholding operation using the HSV colour space is applied to images3 convolutional layers and 3 pooling layers are used in the activation function called RELU is used in [2] between these convolutional and pooling layers

Optical Character Recognition by Open-Source OCR Tool Tesseract: A Case Study: Applied optical character recognition Tesseract(open source OCRengine) an automated system to convert scanned and printed images, handwritten text to editable text, to recognize the vehicle number plates. They compared the performance of tesseract with another OCR Transym. It was concluded that Tesseract was always faster and had better accuracy.

An Overview of the Tesseract OCR Engine: Explains the pipeline of Tesseract, the first step is a connected component analysis where the outlines are gathered by nesting and converting it to a Blob, further Blobs are grouped into text lines, and the lines and regions are processed for fixed pitch or proportional text. Recognition undergoes a two-pass process to recognize each word, word which is recognized well is passed to an adaptive classifier as training data and in the second pass, the entire page is analyzed to recognize the word which was not recognized well in the previous pass. It also mentions adaptive classifier is more apt for OCR than a static classifier used in a tesseract.

Prediction of sign language has to happen in real-time as it involves dynamic gestures (facial, body pose, hand pose). Image capture, processing and real-time decision making of the sign language is a very computationally intensive task. Low latency image

processing and high FPS is significant in improving the efficiency of the recognition of the sign.

Vastly used Procedural image processing algorithm is computationally intensive which leads to frame loss and a decrease in video streaming due to delay in the computation. Introduction of Multithreading computation by taking advantage of the multi-coreprocessor present in most modern embedded devices

[8] is a survey about how to take advantage of the multi-core architecture of the processor without having knowledge of it by using different software systems such as APIs ex. OPENMP, MPI etc. and explores different language support parallelism in programming.

[9] propose PPcM (parallel processing coincident multi-threading) to process multiple framessimultaneously in different threads. They use library imutilsto perform basic multi-threading. This proposed approach increased the FPS and processing time compared to a single thread.

[7] the article describes OPENMP it is an open multicore processing API. It is one of the open-sourceAPIs available for achieving parallel computation in a muti core processor. It discusses loop level parallelism, variable scope and schedule to achieve parallel computationefficient thread-level parallelism (ETLP) scheme is proposed in [10]. They use an edge detection algorithm for evaluation purposes.it utilizes a multi-core processor effectively and decreased the execution time of the algorithm.

III. CONCLUSION

In this paper, we have listed all the techniques that are useful in building a stand-alone Sign Language translator deployed on a Raspberry Pi, that can convert dynamic fingerspells and predict the corresponding words and form a sentence with the help of the Hand-mesh model available in the Mediapipe framework. The generated sentence will be converted into audio form as well. In addition to this, we have employed emotion recognition using the face-mesh model which is also present in MediaPipe. We have also integrated a method that can recognize images of text embedded on surfaces such as boards or flyers etc., This recognized text was successfully translated into a regional language of our choice using Google text-to-speech API.

Sl No.	Title	Method	Advantages	Disadvantages
[1]	Lugaresi, C., Tang, J., Nash, H., Mediapipe: A framework for building perception pipelines	MediaPipe: Pose/ Hand/ Face Recognition	Ready to use Solution, Easy to Integrate and works on Cross-Platform, Open-Sourced	Need to provide good quality images (Lighting Conditions, Contrast and Brightness)
[2]	Smith, R. (2007, September). An overview of the Tesseract OCR engine. In the Ninth international conference on document analysis and recognition	PyTesseract OCR Engine	Faster and more Accurate than other OCR Engines	Prediction is unstable if image contains Noise
[3]	Brahmana, Sofyan, R., & Putri, D. M. (2020). Problems in the Application of Google Translate as a Learning Media in Translation.	Google Translate API	Easy to use and Integrate	The translation is inaccurate and inappropriate sometimes Requires Internet Connectivity
[4]	Georgios Pavlakos, Vasileios Choutas. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image	3D model of human body pose, hand pose, and facial expression	2D features and then optimizes model parameters to fit the features	Number of cameras required is more to train the model(paired images)
[5]	UniPose: Unified Human Pose Estimation in Single Images and Videos	UniPose, a unified framework for human pose,"Waterfall" Atrous Spatial Pooling architecture(WASP)	The architecture used in the model is simple	Only estimates pose of the human body excludes facial expression
[6]	James Jenn-Jier Liena,b, Takeo Kanade. Detection, tracking, and classification of action units in facial expression	Dense-flow extraction using a wavelet motion model, facial-feature tracking, and edge and line extraction	Captures minute details of the facial expression	The complexity of the run-time is more

Table Ⅰ: Comparison table

REFERENCES

[1] Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D. and Black, M.J., 2019. Expressive body capture: 3d hands, face, and body from a single image. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10975-10985).

[2] Smith, R., 2007, September. An overview of the Tesseract OCR engine. In Ninth international conference on document analysis and recognition (ICDAR 2007) (Vol. 2, pp. 629-633). IEEE.

[3] Brahmana, C.R.P.S., Sofyan, R. and Putri, D.M., 2020. Problems in the application of Google Translate as a learning media in translation. Language Literacy: Journal of Linguistics, Literature, and Language Teaching, 4(2), pp.384-389.

[4] Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A., Tzionas, Di., & Black, M. J. (2019). Expressive body capture: 3D hands, face, and body from a single image. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Vol. 2019-June, pp. 10967–10977). IEEE Computer Society.

[5] Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D. and Black, M.J., 2019. Expressive body capture: 3d hands, face, and body from a single image. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10975-10985).

[6] Lien, J.J.J., Kanade, T., Cohn, J.F. and Li, C.C., 2000. Detection, tracking, and classification of action units in facial expression. Robotics and Autonomous Systems, 31(3), pp.131-146.

[7] Slabaugh, G., Boyes, R. and Yang, X., 2010. Multicore image processing with openmp [applications corner]. IEEE Signal Processing Magazine, 27(2), pp.134-138.

[8] Kim, H. and Bond, R., 2009. Multicore software technologies. IEEE Signal Processing Magazine, 26(6), pp.80-89.

[9] Shammi, S.K., Sultana, S., Islam, M.S. and Chakrabarty, A., 2018, June. Low Latency Image Processing of Transportation System Using Parallel Processing co-incident Multithreading (PPcM). In 2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR) (pp. 363-368). IEEE.

[10] Indragandhi, K. and Jawahar, P.K., 2020. An Application based Efficient Thread Level Parallelism Scheme on Heterogeneous Multicore Embedded System for Real Time Image Processing. Scalable Computing: Practice and Experience, 21(1), pp.47-56.

[11] Bagby, B., Gray, D., Hughes, R., Langford, Z. and Stonner, R., 2021. Simplifying sign language detection for smart home devices using google mediapipe.