World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Real-Time Vision-Based Sign Language Bilateral Communication Device for Signers and Non-Signers using Convolutional Neural Network

Uriah Sampaga *, Andrea Louise J. Toledo, Mikayla Assyria L. Dela Peret, Luisito M. Genodiala, Sheika Rania D. Aguilar, Gellie Anne M. Antoja, Charles G. Juarizo and Eufemia A. Garcia

Department of Electronics Engineering, College of Engineering, Pamantasan ng Lungsod ng Maynila, Manila, Philippines.

World Journal of Advanced Research and Reviews, 2023, 18(03), 934–943

Publication history: Received on 07 May 2023; revised on 15 June 2023; accepted on 17 June 2023

Article DOI: https://doi.org/10.30574/wjarr.2023.18.3.1169

Abstract

The use of sign language is an important means of communication for individuals with hearing and speech impairments, but communication barriers can still arise due to differences in grammatical rules across different sign languages. In an effort to address these barriers, this study aimed to develop a real-time two-way communication device that uses image processing and recognition systems to translate two-handed Filipino Sign Language (FSL) gestures and facial expressions into speech; the system can recognize gestures that correspond to specific words and phrases. Specifically, the researchers utilized Convolutional Neural Networks (CNNs) to enhance the processing speed and accuracy of the device. The system also includes a speech-to-text (STT) feature that helps non-signers communicate with deaf individuals without relying on an interpreter. The study's results showed that the device achieved a 93% accuracy rate in recognizing facial expressions and FSL gestures using CNN, indicating that it is highly accurate. Additionally, the system performed in real-time, with an overall average conversion time of 1.84 and 2.74 seconds for sign language to speech and speech to text, respectively. Finally, the device was well-received by both signers and non-signers, with a total approval rating of 85.50% from participants at Manila High School, suggesting that it effectively facilitates two-way communication and has the potential to break down communication barriers.

Keywords: Filipino Sign Language; Two-Way Communication; Facial Expression Recognition; Convolutional Neural Networks

1. Introduction

Communication is a crucial life skill that individuals use every day to convey information through verbal or nonverbal means. In 2015, approximately 70 million people worldwide were deaf, according to statistics from the World Federation of the Deaf (WDF). In his study claims that while individuals with hearing and speech impairments can communicate effectively amongst themselves, challenges arise in educational, social, and work environments due to the lack of knowledge about sign language [1].

Sign language recognition devices such as glove-based approach and vision-based approach were developed to assist people with hearing and speech impairment. [3] [4] indicated that the glove-based sign language recognition system requires expensive hardware and can be uncomfortable for users to wear. The study of [2] [5] related to vision-based sign language recognition systems only exhibiting one-way communication and only capable of determining sign languages that are letters and numbers using only one hand. Moreover, incorporating facial expression recognition and hand gesture recognition can result in a more accurate sign language recognition device [5], however, existing sign language recognition systems lack facial expression recognition features. Therefore, developing a vision-based sign

* Corresponding author: Uriah Sampaga

Copyright © 2023 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution Liscense 4.0.

language two-way communication device with the integration of facial expression recognition is important to bridge the communication gap between signers and non-signers.

Given the importance of utilizing sign language for communication among deaf individuals, this study seeks to develop a real-time vision-based sign language bilateral communication device that will detect hand gestures, facial expressions, and voice that will be converted into its corresponding speech or text output. Specifically, it aims to: (1) develop a successful facial expression and sign language recognition system using the Convolutional Neural Network (CNN); (2) test the real-time system of facial expression and sign language to speech and speech to text conversion; (3) validate the effectiveness of a real-time vision-based sign language bilateral communication device for signers under the special education (SPED) program and non-signer students at Manila High School.

2. Material and methods

2.1. System Architecture

The key components of the two-way communication device were the web camera, Python program, database, interpretation, integration, and conversion. A standalone local network solution, Python, and information from the web camera were used to create the software architecture for the sign language recognition systems. Python libraries and input from the microphone were used to construct the software architecture for speech recognition and text conversion. Figure 2 illustrates a system consisting of a web camera, Python software, speaker, microphone, and monitor. The web camera captures sign language and facial expressions, converting the analog data to digital using Python's analog-to-digital converter. The captured image sequence is continuously collected, interpreted, and saved for processing by the Python software. The speaker outputs speech based on the integrated data from Python, while the microphone receives speech input from the non-signer, which Python processes to convert into text. The monitor displays the text output for the signer. Python is crucial in hand detection, hand tracking, face detection, image processing, sign language recognition and translation, and speech to text conversion. The Python Software utilizes OpenCV for hand and face detection, Keras and Tensorflow for real-time sign language recognition using a CNN model, and PyAudio and SpeechRecognition for speech to text conversion. The system stores collected data in a repository for scalability and updates.

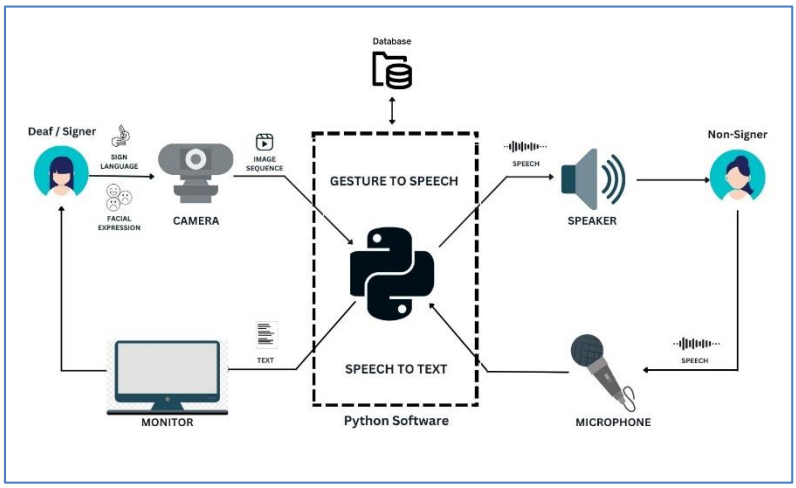

Figure 1 System Architecture

2.2. Model Training

2.2.1. Dataset Acquisition

The dataset for training the static, dynamic, and facial expression sign language models was gathered by manually capturing images using Python and OpenCV libraries. Images underwent a series of image processing procedures such as hand detection, facial detection, landmarks application, cropping, and overlaying the images on a plain white background. The Hand Detector Module and Haar Cascade Classifier were responsible for the detection process. Images are saved to their respective folder in the local repository using the os library.

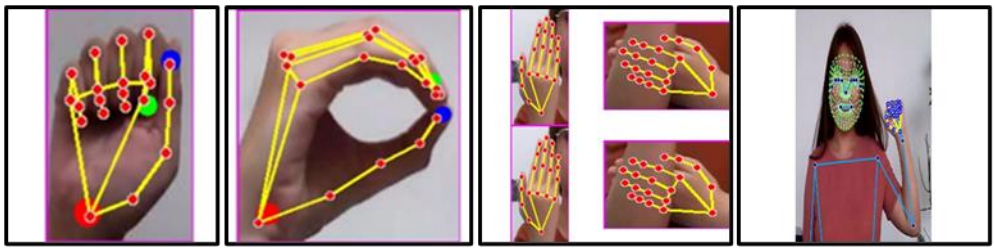

Figure 2 Sample of Sign Language and Facial Expression Datasets

2.2.2. Network Layers

The researchers used transfer learning to create each recognition system's Convolutional Neural Network models. The MobileNetV2 pre-trained model was used as a based model followed by a GlobalAveragePooling2D layer, a dense layer with 1000 neurons, and a 50% dropout layer. Another dense layer was added, followed by a SoftMax activation function to complete the optimized CNN architecture.

2.2.3. Training

The model was trained through Keras library and TensorFlow as its backend using a laptop with a Graphics Processing Unit GeForce RTX 3050Ti. Adam optimizer was used as an optimizer with a learning rate of 0.001 to compile the network. Moreover, categorical cross-entropy was used as the loss function. The default epoch was set to 50 with a batch size of 16. The training process utilized the ModelCheckpoint and EarlyStopping modules to monitor the process and prevent the model from overtraining.

2.3. Testing Methodology

The testing methodology comprises different tests for evaluating the device recognition and conversion accuracy, speed, and effectiveness. The device was tested by four (4) Junior High School and Senior High School students, two (2) are from the Special Education program, and two (2) are from the Regular program at Manila High School.

The sign language recognition system, sign language to speech system, and speech to text system testing procedure was comprised of ten (10) trials. Each trial had a duration of five (5) seconds. The time it took for the systems to produce their respective outputs was monitored using a stopwatch. The device was also assessed using a Likert scale with a five-point rating to validate the effectiveness of the real-time vision-based sign language bilateral communication device.

2.3.1. Testing Procedure

Before initiating any sign language recognition system, the signers must stand in front of the device at a distance of one foot. The device required the signers to press A, N, D, or F to start the static, dynamic, or facial sign language recognition system. The signers were tasked to perform and repeat each sign language ten (10) times. The output was only considered correct if the system could produce the equivalent speech output of the signs correctly within five (5) seconds. The researchers determined if there is a significant difference between the means of training accuracy rate and testing accuracy rate of the recognition systems using the independent two-sample t-test as a statistical method.

The speed of the sign language to speech system was evaluated by having the signers perform the sign language and press the "o" button to produce the equivalent speech output of the letter, number, word, and phrase. For the speech to text system, the non-signer must stand in front of the device, not one foot away, and press the "0" button that initiated the system. The non-signer was tasked to press the space bar button, release it after 3 seconds, and then speak the predetermined greetings and questions. The procedure for both systems was repeated ten (10) times. The outputs were only considered correct if the system could produce the equivalent text output within five (5) seconds. The researchers recorded the number of correctly and incorrectly displayed sentences to obtain the percentage accuracy of the system. The recorded time was used to evaluate the real-time conversion capability of the sign language to speech system and speech to text system.

Each student was handed a survey questionnaire with ten (10) questions regarding the device's functionality, dependability, usability, efficiency, and learning impact. The survey questionnaire utilized the Likert scale with a five-point rating which consists of the responses excellent, good, fair, average, and poor; excellent = 5, good = 4, fair = 3,

average = 2, poor = 1. The total score was calculated by adding the user's ratings in questions about each criterion. The total score validated the effectiveness of a real-time vision-based sign language bilateral communication device.

2.3.2. Percentage Accuracy formula

The percentage accuracy of the systems was determined by getting the sum of the correctly recognized letters, numbers, words, phrases, or displayed sentences and dividing it by the product of the total number of participants multiplied by the number of trials.

The output was considered correct recognition when the system translated the sign language, facial expression, and speech to their respective equivalents within five (5) seconds. It is considered incorrect if the system produces a different output outside the allocated time.

$$\%\ Accuracy = \frac{Total\ Nos.of\ Correct\ Recogniton/Display}{(Total\ no.of\ Users)(No.of\ Trials)} \quad (1)$$

2.3.3. Average Time (s)

The number of seconds it took for each system to produce a correct output in each trial conducted was utilized to assess the learning accuracy of the system. Moreover, the results from each trial were also averaged.

$$Average\ Time(s) = \frac{\Sigma x}{n} \quad (2)$$

where:

x: sum of results of each trial
n: total number of trials

2.3.4. Approval Percentage Formula

The approval percentage is obtained by dividing the total score by the goal score.

$$Approval\ Percentage = \frac{Total\ Score}{Goal\ Score} \times 100 \quad (3)$$

2.3.5. Independent Two-Sample T-Test

The significant difference of the training accuracy rate and testing accuracy rate was determined and analyzed using the independent two sample T-Test. The independent two-sample t-test is a statistical test used to compare the means of two independent samples. If the t Stat value is more significant than the t Critical two-tail or t Stat is less than the negative of t Critical two-tail, and the p-value is less than the predetermined significance level (0.05), the null hypothesis would not be accepted.

3. Results and Discussion

The accuracy percentage and conversion speed are crucial as they will impact the effectiveness and reliability of the real-time vision-based sign language bilateral communication device. The percentage accuracy of each test is calculated using formula (1) and formula (2) for the average time. Each test is composed of ten (10) trials.

Table 1 shows the results of the facial expression recognition accuracy test and facial expression to speech conversion speed test. The words ‘Yes' and 'Sorry' were observed to have an accuracy rating of 100% from both signer students. On the other hand, there are some incorrectly recognized words, 'Don't understand,' 'No,' and 'Please,' which result in an accuracy rating of 85%; this implies that the system had difficulties detecting these words. Moreover, the facial expression for ‘Yes’ into speech was 2.01 seconds suggesting that ‘Yes’ is a relatively easy word to convert from facial expression to speech for the system. Similarly, ‘No’ also had a short average time of 1.84 seconds, indicating that the system could quickly convert this facial expression into speech.

The total accuracy rating for facial expression recognition was 91%, and the average time of the facial expression to speech conversion system was 2.03, which indicated that the facial expression to speech system was highly effective in recognizing that most sign languages accurately convert facial expressions into speech in real time.

Table 1 Facial Expression Recognition System Accuracy and Facial Expression to Speech Conversion Speed Test

Words	Correctly Recognized Gestures	Incorrect Recognized Gestures	Accuracy (%)	Average time (s)
Don't Understand	17	3	85	2.27
Yes	20	0	100	2.01
No	17	3	85	1.84
Please	17	3	85	2.03
Sorry	20	0	100	1.98
	Overall Rating		91%	2.03

Table 2 Letter Recognition System Accuracy and Letter to Speech Conversion Speed Test

Letter	Correctly Recognized Gestures	Incorrect Recognized Gestures	Accuracy (%)	Average time (s)
A	20	0	100	1.6
B	20	0	100	2.75
C	20	0	100	2.02
D	20	0	100	1.86
E	20	0	100	2.09
F	20	0	100	2.12
G	17	3	85	1.52
H	20	0	100	1.45
I	20	0	100	1.78
J	14	6	70	2.29
K	17	3	85	1.86
L	20	0	100	1.56
M	12	8	60	2.13
N	17	3	85	1.9
O	20	0	100	1.23
P	18	2	90	1.65
Q	17	3	85	1.56
R	20	0	100	1.34
S	20	0	100	1.43
T	20	0	100	1.6
U	17	3	85	1.29

V	20	0	100	1.19
W	20	0	100	1.45
X	20	0	100	1.59
Y	20	0	100	1.56
Z	17	3	85	1.05
	Overall Rating		93.46%	1.69

Table 2 depicts the results of the letter recognition accuracy test and letters to speech conversion speed test. The percentage accuracy rate was 100% for 17 specific alphabets, namely A, B, C, D, E, F, H, I, L, O, R, S, T, V, W, X, and Y. The letter M obtained the lowest percentage accuracy rate of 60%, indicating that this letter could have been more challenging for the system to recognize. Additionally, letter Z had the fastest average time, and letter B had the slowest average time. It indicated that converting the letter B into speech was more challenging. This could be due to the visual complexity of this letter, making it more difficult for the system to convert it into speech quickly.

The overall speed for letter conversion acquired a total of 1.69 seconds, and the total accuracy rating for letter recognition was 93.46%. The results showed that the sign language recognition system was highly effective in recognizing most letters and can convert sign language in the form of letters to speech in real time.

Table 3 Number Recognition System Accuracy and Number to Speech Conversion Speed Test

Number	**Correctly Recognized Gestures**	**Incorrect Recognized Gestures**	**Accuracy (%)**	**Average time (s)**
0	20	0	100	1.31
1	20	0	100	1.51
2	20	0	100	1.75
3	14	6	70	2.11
4	20	0	100	1.98
5	20	0	100	1.41
6	14	6	70	2.16
7	20	0	100	1.73
8	17	3	85	1.7
9	17	3	85	1.33
		Overall Rating	91%	1.70

Table 3 demonstrated that the sign language recognition system was 100% accurate for most numbers (0, 1, 2, 4, 5, 7). The results from the speed test also indicated that the average conversion times only ranged from 1.31 seconds up to 2.16 seconds. The system converts most numbers (0, 1, 2, 4, 5, 7, 8, 9) below two (2) seconds.

The total accuracy rating for number recognition was 91%, with an overall speed of 1.70 seconds. The test results suggested that the numbers sign language recognition and conversion system could efficiently and accurately recognize and convert most numbers into speech in real time.

Table 4 Greetings Recognition System Accuracy and Greetings to Speech Conversion Speed Test

Greetings	Correctly Recognized Gestures	Incorrect Recognized Gestures	Accuracy (%)	Average time (s)
Hello	17	3	85	1.69
Good morning	20	0	100	1.99
Good afternoon	20	0	100	2.30
See you tomorrow	20	0	100	2.57
Thank you	14	6	70	1.96
Welcome	17	3	100	2.5
		Overall Rating	92.50%	2.09

Table 4 shows the accuracy percentage of each greeting, where most results are 100% accurate. However, some incorrect gestures, like 'Hello' and 'Thank you,' resulted in a lower accuracy of 85% and 70%, respectively. The speed test results indicate that all greetings had an average conversion time of below three (3) seconds and ranged from 1.69 seconds to 2.57 seconds. 'See you tomorrow' had the slowest average time of 2.57 seconds, indicating that converting this greeting was more challenging because it contains more words and is, therefore, more complex than other greetings.

The overall accuracy rating of greetings recognition obtained a 92.50% accuracy rate with an overall speed of 2.09 seconds. The results show that the greetings recognition system was considered real-time since all the greetings were converted into speed within five (5) seconds.

Table 5 Answer Recognition System Accuracy and Answer to Speech Conversion Speed Test

Answers	Correctly Recognized Gestures	Incorrect Recognized Gestures	Accuracy (%)	Average time (s)
Letters	20	0	100	1.71
Yes/No	20	0	100	2.04
Numbers	20	0	100	1.91
Time	17	3	85	1.71
Yes/No	20	0	100	2.16
		Overall Rating	97%	1.90

Table 5 depicts that the sign language recognition systems used for answering the non-signers' questions had excellent results that resulted in a 100% accuracy rate of each answer from Sign Language to Speech, except for 'Time,' resulting in a lower accuracy of 85%. Moreover, speed test results indicate that all answers had an average conversion time of below three (3) seconds. The 'Letters' and 'Time' were the quickest to be converted, taking an average time of 1.71 seconds.

Overall, findings suggest the system is relatively fast and accurate, having an average 1.90 seconds and an accuracy percentage of 97%.

Table 6 Greetings Speech to Text Conversion Accuracy and Speed Test

Greetings	Correctly Recognized Gestures	Incorrect Recognized Gestures	Accuracy (%)	Average time (s)
Hello	20	0	100	2.43
Good morning	20	0	100	2.55
Good afternoon	20	0	100	2.75
See you tomorrow	20	0	100	2.99
Thank You	20	0	100	2.91
Welcome	20	0	100	2.59
	Overall Rating		100%	2.09

Table 6 Shows that "Hello" has the shortest display speed of 2.43 seconds while "See you tomorrow" has the most prolonged display of 2.99 seconds. This implies that the shorter greetings displayed faster to the monitor from speech to text. The accuracy test also showed that the speech to text conversion system achieved an overall rating of 100% in all trials for all tested greetings. The display of greetings was completed with an overall speed of 2.70 seconds with 100% accuracy, which suggests that the average time for displaying the words/phrases for greetings was within a five (5) second duration and was considered accurate and can convert speech in real time.

Table 7 Questions Speech to Text Conversion Accuracy and Speed Test

Questions	Correctly Recognized Gestures	Incorrect Recognized Gestures	Accuracy (%)	Average time (s)
What's your name?	20	0	100	2.31
Are you new to this school?	17	3	85	3.033
How many years have you been studying here?	20	0	100	3.085
What time does your school start?	17	3	85	3.145
Do you understand?	20	0	100	2.33
	Overall Rating		94%	2.78

Table 7 shows that the speech to text conversion system was able to accurately display all five (5) questions to text with a percentage accuracy of 100% for "What's your name?", "How many years have you been studying here?", and "Do you understand?". However, the questions "Are you new to this school?" and "What time does your school start?" obtained a percentage accuracy of 85%. The speed test for the conversion system showed that "What time does your school start?" has the longest average time of 3.15 seconds and "What's your name?", which is relatively shorter than other questions and has an average display speed of 2.31 seconds. The display of questions was completed with an overall speed of 2.78 and an accuracy percentage of f 94%. This indicated that the speech to text conversion system could accurately display the questions in real time. In addition, variations in time for each question indicate that the time taken depends on the length and complexity of the phrase or sentence.

Table 8 Independent Two-Sample T-Test Results

	Z	Training Accuracy vs Testing Accuracy
Words	t Statistic	2.449489743
	P (T ≤ t), two-tailed	0.039968524
	t Critical, two-tailed	2.008559112
Letters	t Statistic	3.128893239
	P (T ≤ t), two-tailed	0.002925877
	t Critical, two-tailed	2.008559112
Numbers	t Statistic	2.117294717
	P(T ≤ t), two-tailed	0.048419178
	t Critical, two-tailed	2.1009220
Greetings	t Statistic	1.812461123
	P(T ≤ t), two-tailed	0.03697419
	t Critical, two-tailed	1.463850109

In order to further evaluate the testing accuracy of the sign language recognition systems, a comparison between the mean of correctly and incorrectly recognized gestures in each trial, and the mean of the training accuracy was conducted. This analysis was performed using the Independent Two-Sample T-Test at a significance level of 0.05. The summary of the t-test results is presented in Table 8.

The null hypothesis is that the means of the groups are equal. To fully reject the null hypothesis, both p-values and t-values were examined. The obtained p-values for the samples were below the significance level (0.05), leading to the rejection of the null hypothesis. Additionally, the t Statistic should be greater than the t Critical. Table 8 presents results indicating that the t Statistic values for words, letters, numbers, and greetings were greater than the t Critical values, thereby demonstrating a significant difference between the means of each group. This substantial evidence supports the full rejection of the null hypothesis.

Table 9 Summary of Evaluation Results

	No. of Evaluator	No. of Questions	Total Score	Goal Score	Approval Percentage
Functionality	4	2	35	40	87.50%
Reliability	4	2	33	40	82.5%
Usability	4	2	32	40	80%
Efficiency	4	2	35	40	87.50%
Learning Impact	4	2	36	40	90%
			Total Approval Percentage		85.5%

The evaluation results of the real-time vision-based sign language bilateral communication device, summarized in Table 9, demonstrated its effectiveness across multiple criteria. The device received high approval percentages for learning impact (90%), functionality and efficiency (87.50%), reliability (82.50%), and usability (80%). These findings validated the device's ability to provide a positive learning experience, perform its intended functions effectively and efficiently, and offer ease of use and reliability. With an overall approval percentage of 85.50%, the evaluation results verified the device's effectiveness as a reliable tool for facilitating two-way communication in sign language.

4. Conclusion

The study successfully developed a facial expression recognition system, sign language to speech system, and speech to text system, achieving high accuracy rates and efficient real-time conversion. The vision-based sign language bilateral communication device received positive ratings. This study will contribute to bridging the communication gap between signers and non-signers, enabling effective and real-time communication using sign language and speech. Further advancements can be explored to enhance its capabilities and expand its application.

Compliance with ethical standards

Acknowledgements

The researchers would like to thank the Division of City Schools Manila, as well as Manila High School and Pamantasan ng Lungsod ng Maynila for extending their profound knowledge and for the endless support towards the realization of this study. The researchers would also like to acknowledge the Department of Science and Technology (DOST), the funding agency, for supporting the research. Additionally, the researchers would like to express their gratitude to their advisers, Engr. Charles G. Juarizo and Engr. Thaddeo S. Garcia; and to the panelists, Engr. Eufemia A. Garcia, Engr. Fernando Victor V. de Vera, and Prof. Mark Christopher R. Blanco for granting them the opportunity, offering guidance and being a source of inspiration during this research endeavor.

Disclosure of conflict of interest

The authors of this manuscript declare that there are no conflicts of interest in any form.

Statement of Ethical Approval

This statement certifies that the appropriate review board has granted ethical approval to the research study titled Real-Time Vision-Based Sign Language Bilateral Communication Device for Signers and Non-Signers using Convolutional Neural Network. The study adheres to ethical standards, assuring participant well-being, informed permission, and anonymity. This clearance confirms the project's adherence to ethical standards and commitment to ethical research.

Statement of Informed Consent

Informed consent was obtained individually by agreeing to be part of the study.

References

[1] Alam RF, Bin Munir M, Ishrak S, Hussain S. A Machine Learning Based Sign Language Interpretation System for Communication with Deaf-mute People [Internet]. 21th International Conference on Human Computer Interaction; 2021. Available from: https://www.researchgate.net/publication/353018855_A_Machine_Learning_Based_Sign_Language_Interpretation_System_for_Communication_with_Deaf-mute_People

[2] Anudeep, K., Goud, P., Harsha, K., & Swamy, K. (2018, February). Sign to Speech and Display Converter. Vol. 1, Issue 8, pp 85-89. https://www.irejournals.com/formatedpaper/1700216.pdf

[3] Hurro M, Elham M. Sign Language Recognition System using Convolutional Neural Network and Computer Vision [Internet]. 2020. Vol. 9 Issue 12, pp. 59- 64. Available from: https://www.ijert.org/sign-language-recognition-system-using-convolutional-neural-network-and-computer-vision

[4] Shah S, Nisar K, Udeshi AJ, Kotia A. A Vision Based Hand Gesture Recognition System using Convolutional Neural Networks [Internet]. International Research Journal of Engineering and Technology (IRJET).; 2019. Vol. 6 Issue 4, pp. 1-6. Available from: https://www.researchgate.net/publication/337274997_A_Vision_Based_Hand_Gesture_Recognition_System_using_Convolutional_Neural_Networks

[5] Sharma A, Panda S, Verma S. Sign language to speech translation [Internet]. 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT).; 2020. Available from: https://ieeexplore.ieee.org/document/9225422