International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

D&D CONVEYOR

Kirithika S[1], Perarasu S[2], Gopinathan S[3], Sathya Narayanan P[4], Ganga M[5], Vigneshwar S I[6]

[1,3,5]Assistant professor
[1,2,3,4,5,6]Department of Artificial Intelligence and Data Science,
Sri Sai Ram Institute of Technology,
Chennai, Tamil Nadu, India.

Abstract
Living in the world with normal people and dumb people are not able to communicate properly. And it's important to provide a place to communicate because deaf and mute people use sign language for communication but they find difficulty in communicating with normal people who don't understand sign language. They also facing many problems like to communicate their feelings to us. To overcome this problem, there is a need of translator to understand what they speak and communicate with us. Computer vision processes the live video into images and the pictures of hand gesture are processed by CNN (Convolutional neural network) in Deep learning. The sign language translation system translates the normal sign language to speech and hence makes the communication between normal person and mute people easier. So, the whole idea is to build a communication system that enables communications between speech hearing impaired and a normal person.

Keywords – Mute people, Deep learning, Hand gestures recognition, Convolutional neural network, Computer vision.

1. INTRODUCTION

Communication is vital in the modern world in terms of interaction, collaboration, and inclusion. But the deaf and the mute or people with hearing and speech disability have a big problem in an environment where spoken language is the most common form of communication because of their inability to communicate and comprehend others. Technology can therefore be used to fill this gap by providing creative solutions that would ensure that these individuals can communicate straight away. The proposed project aims at creating a Deaf and Dumb Communicator based on Computer Vision (CV) and Deep Learning (DL) methodologies. This system was mainly aimed at allowing real time identification and translation of sign language gestures to text or speech that would simplify communication among the people with hearing and speech difficulties. This project will use the development of image processing, gesture recognition and neural networks to develop an intuitive and efficient tool that will bring greater accessibility and social inclusion.

The system records the hand gestures by a camera, analyzes them with the help of deep learning models and turns them into significant text or speech. Computer vision enables tracking of any movement of hands with high accuracy, but deep learning models enable a good recognition of the movements, even when it comes to the complex and dynamic gestures. This project does not only help

with enhancing communication among people with disabilities but also becomes a step into the limits of AI and machine learning applications in the real world.

To sum up, this project proposes a technologically instigated solution to communication barriers among the deaf and mute community to facilitate a less exclusive and more socialized society.

2. LITERATURE REVIEW

Gesture recognition and sign language have been given a lot of attention, because it is essential in the field of assistive technology and human computer interaction. The review of sign language recognition systems provided by Rojas and Orozco [1] is broad, dividing them into the vision-based and sensor-based types. Their analysis emphasizes the shift of the hand-made aspects to the machine-learning-based approaches and reveals the key challenges that include signer dependency, the complexity of background, and scanty annotated datasets. Yu and Miao [2] concentrate on the deep learning methods of gesture recognition and show that convolutional and recurrent neural networks contribute a great deal to the spatial-temporal features of gestures extraction. Nevertheless, they observe that the cost of high computation and data dependency is still a significant hinder. Alkhazaleh and Ali [4] suggest a real-time sign language recognition system through deep learning methods, which can be enhanced with better accuracy and lower latency, but intense preprocessing is necessary to address the real-life situations. Zha and Wu [6] offer a universal overview of sign language recognition, and examine datasets, evaluation procedures, and model structures; the inability to unify benchmarking and the scope of generalization are identified as gaps in the research. Mirończuk and Pękala [9] continue to overview state-of-the-art gesture recognition systems in various sensing modalities and conclude that cross-domain adaptability and robustness remains an issue that is not completely addressed.

Multimodal interaction systems are incomplete without facial emotion recognition. Kafai and Makarova [3] provide a review of deep-learning-face emotion recognition along with demonstrating that convolutional neural networks are more effective compared to conventional techniques. However, they note that their review has identified ongoing problems like facial occlusion, differences in illumination and cultural bias in emotion datasets. Poria and Mihalcea [7] prove the effectiveness of the multimodal feature fusion by showing that multimodal emotion recognition based on visual, audio, and textual features is more accurate than unimodal system. Speech processing technology is used together with gesture and emotion recognition in multimodal systems. This publication by Hinton et al. [8] provides the background of the deep neural network in acoustic modeling of speech recognition in proving to be highly superior to the traditional Gaussian mixture models and laying the foundation of the contemporary speech recognition systems. The article by Ganaie and Shah [5] embraces the latest trends in text-to-speech synthesis, presenting the development of neural TTS systems, which enhance the naturalness of speech and the ability to control the prosody. They, however, mention difficulties associated with low-resource languages as well as measures of evaluation.

The intention of multimodal communication systems is to combine speech, gesture, and emotion so that natural and effective interaction can be achieved. D'Mello and Graesser[5] discuss multimodal communication models and demonstrate that using both verbal and non-verbal messages enhances better

understanding of the system and involvement of users. Their contribution is that proper integration and coordination of modalities are the key to the successful multimodal interaction.

All in all, the literature illustrates that there has been a substantial advancement in sign language recognition, analysis of gesture, emotion analysis, and speech technologies based on deep learning and multimodal fusion. In spite of these developments, issues like the size of datasets, unstandardized assessment, computational complexity, and practical robustness are all research issues. These problems need to be addressed in order to come up with scalable and dependable multimodal human-computer interaction systems.

3. PROBLEM STATEMENT

The main issue, which people with hearing and speech impairments can encounter, is that it is hard to effectively communicate with the rest of the population in the spoken language setting.

Although the use of sign language has been wide among the deaf and mute community, majority of the non-impaired people are not conversant with this language, posing a communication barrier. This inability to understand each other makes people with disabilities unable to engage in social, educational, and professional processes fully. This is further worsened by the fact that there are no cheap real- time translation applications that can help fill this gap.

Hence, there is an urgent requirement of a system that can translate sign language to either text or speech in real time to enable communication between the hearing impaired, speech impaired and ordinary people. The proposed project will help mitigate this issue by creating a sign language translation device named Deaf and Dumb Communicator, which will recognize and decode sign language gestures automatically to decipher them into a significant language through the use of computer vision and deep learning methods.

4. RELATED WORKS

Many researches have explored the use of computer vision and machine learning in the creation of systems capable of detecting sign language and helping the hearing or speech- impaired population. Initial solutions were based on the wearable equipment such as data gloves or motion sensors to record the hand gesture. But with the evolution of computer vision, the camera based methods became popular owing to the lack of intrusion. Some of the machine learning models that have been used in gesture recognition include Support Vector Machines (SVM) and Hidden Markov Models (HMM). Although these conventional techniques proved to be successful in detecting the non-dynamic gestures, they could not detect continuous dynamic gestures that are frequently employed in sign language. Over the last few years, deep learning has become a better tool to deal with the complexity of sign language recognition. The algorithms typically employed are the Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to improve the recognition accuracy of the static and dynamic gestures. The models have realized considerable enhancement in the identification of continuous sign language by combining spatial feature extraction and temporal sequence processing.

Also, real-time sign language recognition, which runs on algorithms such as the YOLO (You Only Look Once) algorithm, has been created to offer real-time translations of hand gestures. Irrespective of these developments, issues like fluctuating lights, noise in the background as well as scalability of such systems are still an ongoing research topic.

5. DATAACQUISITION

To come up with an efficient system of Deaf and Dumb Communicator, computer vision and deep learning are essential, and a large and high-quality dataset is essential. The data acquisition is carried out through capturing a variety of sign language gestures in order to train and verify the model to be precise in recognizing gestures. The data acquisition plan of this project follows:

1. Dataset Selection or Creation: This is where one selects an existing dataset or develops a new dataset. A number of publicly available datasets like the American Sign Language (ASL) dataset can be used by starting to train a model. They are frequently available in the form of images or even videos of static gestures (alphabet signs) and, in some cases, dynamic gestures (words or phrases). Nonetheless, in case the system has to support a particular sign language or an individual set of gestures, then it is required to develop a custom dataset. This would include filming video or series of images of people with different gestures in sign language.

2. Data Collection Process: In case of a custom dataset, collection process would involve capturing hand gestures to angles, and environments to make them robust. One should ensure to collect information of a number of people of different age, gender, and hand size to enhance the diversity of the data. It is advisable to record the gestures in a bright room with a neutral background to reduce these noise and occlusions. The gestures should be photographed in different angles and distance in order to recreate the conditions of use in the real world.

3. Annotation: After data collection, it should be annotated,
   i.e. every frame of an image or a video should be labeled with the specific gesture. In the case of static gestures, the labelling procedure can be characterized by attaching the right label (e.g. words, letters) to each image. In the case of dynamic gestures, the video frames should be marked with the action or sign. Image dataset annotation can be done with tools such as LabelImg or VIA, whereas video annotation can be done with video annotation tools.

4. Preprocessing: Once annotated, the data is supposed to be preprocessed to make it better in terms of quality and usability in training a model. It includes downsizing of images or frames to a uniform resolution, normalizing pixel values and augmentation, such as rotation, flipping or changing brightness of images. Such techniques make the data more diverse to assist the model to be more general to unknown data. Also, the background noise can be reduced by using background subtraction or background segmentation methods, and the model can be dedicated to the hand and gesture areas only.

6. PREPROCESSING TECHNIQUES

Before feeding the images of the captured gestures into the CNN model, preprocessing is required. It makes sure the model gets a clean, coherent and meaningful input which significantly enhances recognition accuracy.

Hand Segmentation

The system divides the hand area and the background first with the help of such methods as thresholding, skin-color detection, or background subtraction. Through isolation of the hand only, the model helps concentrate on the significant shape of gestures rather than on noise surrounding it. This minimizes the mistakes and enhances feature extraction. Image Resizing and Image Normalization

Gesture frames are all down sampled to a constant resolution (e.g.64x64) to provide the model with equal input. The pixel values are then brought to a standard range (0-1) and this assists the CNN to learn patterns with ease and training is made faster.

Data Augmentation

In order to replicate the reality, the data is augmented with the help of manipulations such as rotation, flipping, brightness changes, and zooming. This would train the model to perform the gestures in spite of the lighting variations, even when the hand leans a bit, or the user is in a new position.

Noise Reduction

Visual noise is removed using filters like the Gaussian blur or the median filter. This sharpens the edges and shape of the gesture and the CNN is then able to extract the accurate features of each frame.

7. SYSTEM REQUIREMENTS

In order to have smooth operations of the proposed Deaf and Dumb Communicator system, hardware and software should get the appropriate configuration.

1. **Hardware Requirements**

- Camera: It will need a high-resolution web camera that will record accurate and clear hand gestures.
- Microphone: This device is used to capture the speech input which can then be converted to gestures.
- Processing Unit: A multi-core CPU computer; it is suggested to have a GPU to perform deep learning inference at a faster speed.
- Memory and storage: Memory of 8GB and more and storage of datasets and model files.

2. Software Requirements

- Programming Language: Python to develop the recognition pipeline.
- Libraries: OpenCV (video processing),TensorFlow/Keras (deep learning) NumPy and Pandas (data handling).
- Speech: Speech Recognition library which does audio transcription and pyttsx3 or gTTS which does text-to-speech.
- Operating System: Windows, Linux or MacOS with compatible drivers to camera and microphone.

3. Dataset Requirements

•Supervised learning Annotated gesture images or frames.
•Annotated gesture images or frames for supervised learning.
•Class balance enough to ensure model reliability.

8. METHODS AND MATERIALS

1. EXISTING METHOD

The existing systems designed to support people in communicating with hearing and speech disorder are specifically geared towards recognition of sign language using different technologies. The first systems tended to use hardware based methods, e.g. data gloves with motion sensors, which hear movement of hands and gestures. Although these systems were pretty successful in picking gestures, they were bulky, costly, and were not readily available. The newer systems with the development of computer vision and machine learning have evolved to camera-based solutions that make use of image and video processing to identify hand gestures without wearing extra wearable devices. Classification of static gesture has been done by use of traditional machine learning algorithms such as Support Vector Machines (SVMs) and Hidden Markov Models (HMMs) but they have been found to struggle in real time performance and continual gesture recognition In the recent years, deep learning models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been found to be very successful in improving the accuracy and the efficiency of gesture recognition systems. Nevertheless, some of the current solutions have issues like background noise and changing lighting conditions, as well as the possibility to detect complex and dynamic gestures in real-time.

Fig 1. Current Hand gesture recognition technique.

2. PROPOSED METHOD

2.1 Gesture Recognition to Text and Speech:

- The system involves recording the sign language movements with a camera and interpreting them in the real-time with the use of computer vision and deep learning algorithms.

2.2 Gestures which are identified will be read in form of text which will be written on the screen and then changed into a voice communication by using a text to speech module.

2.3 Facial Emotion Detection:

- The system will recognize the facial emotions of the user like happiness, sadness, or surprise together with the process of gesture recognition.
- Emotion recognition makes communication more interesting and expressive by adding more context to the message and turns communications more natural and expressive.

2.4 Real-time Multimodal Communication:

- The system combines gesture recognition with emotion detection of the face to provide a better, more interactive communication system.
- The system is effective in enabling more meaningful and efficient interactions between the hearing impaired and speech impaired and other people since it gives output in text as well as voice with the emotion sensing.

2.5 Seamless, Accessible Tool:

•The solution offers a real time/easy to use tool that fosters accessibility and which is useful in bridging the communication divide between the hearing and speech impaired.

9. MODEL TRAINING AND HYPERPARAMETERS

The gesture recognition system uses a deep learning model, which is trained using the optimized settings to give good classification.

1. Preparation of Training Dataset

- Dataset will be split into training, validation and testing datasets.
- Labeling of every gesture image is done to facilitate supervised learning.

2. Hyperparameter Configuration

- Learning Rate: This is the rate at which the model is changed during training.
- Batch Size: This is the number of images that are used at any given time during an iteration.
- Epochs: Summation over all the data.
- Optimizer: Adam or SGD parameters to reduce the loss function.
- Loss Function: Categorical cross-entropy loss (Multi- class classification).

3. Model Architecture Settings

- Spatial gesture features are obtained in convolutional layers.
- Pooling layers minimize the dimensions and maintain significant patterns.
- The extracted features are mapped to gesture classes by dense layers.
- Activation is similar to ReLU in that it brings non- linear learning.

4. Performance Monitoring

•Accuracy / loss of training and validation are monitored on a per-epoch basis.
•The preventive measures of overfitting are early stopping or regularization.
•Final model is tested on unknown data of the real world.

10. ARCHITECTURE DIAGRAM

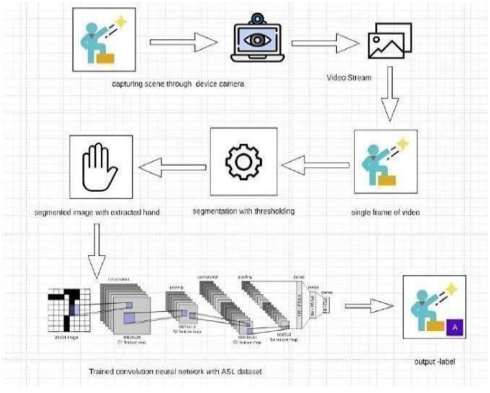

Fig 2. Deaf and Dumb Conveyor Architecture Diagram.

The following architecture figure provides the workflow of a sign language recognition system via a

device camera, which shows the major steps of transforming a video input to the final output label.

1. Input Capture:

The camera of the device records the scene, and this process is based on video streaming in real time. This video feed is the main source of information to identify sign language gestures.

2. Segmentation Process:

The video stream which is captured goes through a process of segmentation which isolates the relevant hand gestures in the background. This will include use of thresholding techniques to isolate the hand in the segmented image, by concentrating on the features which are relevant in the perception of gestures.

3. Single Frame Extraction:

Only one frame is picked out of the segmented video stream and processed further. The isolated hand gesture is found in this frame and this gesture is essential in the following stage of the recognition process.

4. A CNN can be defined in the following way:

The singled out frame is then fed into a trained Convolutional Neural Network (CNN). This was a previously trained model with an American Sign Language (ASL) dataset, which analyzes the image to detect the particular gesture that is being executed. CNN breaks down the image in several layers where each layer is intended to identify various features and patterns which are attributed to hand motions.

5. Output Generation:

Once it has been processed, CNN gives a label, which is the gesture that was recognized. This is the sign that is identified and this sign might be either a letter or a word in the ASL system.

Altogether, the diagram shows a logical process of converting sign language into recognizing the study of the principal image processing and deep learning algorithms and representing real- time video information into actionable outputs. The architecture highlights the significance of every stage in the process of ensuring proper gesture recognition that is the way to effective communication tools to people with hearing and speech disorder.

11. BLOCK DIAGRAM

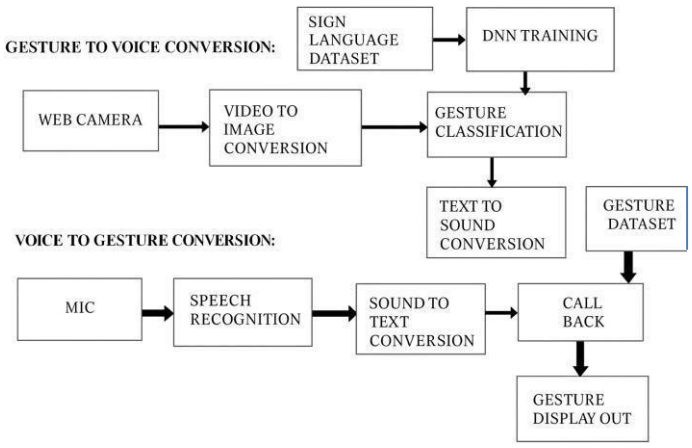

Fig 3. Block Diagram

The block diagram shows the entire process of a two-way communication system that aims at filling the gap between people with hearing or speech impairment and those who can communicate using spoken language. The architecture consists of 2 big modules which are Gesture-to-Voice Conversion and Voice-to-Gesture Conversion which collaborate to facilitate natural meaningful interaction.

1. Gesture-to-Voice Conversion

This module is aimed at the translation of sign language gestures recorded by a camera- to the verbal output.

a. Web Camera Input

A live video feed of a camera is used to start the process. This video records the hands movements continuously of the user, which is the main input in the identification of the gestures of sign language.

b. Video-to-Image Conversion

The video stream is divided into single frame images. These frames simplify the process of the system analyzing the shape, position, and movement of hands.

c. Gesture Classification (DNN-based)

A deep neural network (DNN) is utilized on each extracted image based on a sign language dataset. The model determines a gesture that is being executed based on the learning patterns of the hand structure and orientation. This process is important in that it transforms the raw visual images into a significant

symbolic expression (such as a letter, word or command).

d. Text-to-Sound Conversion

Upon gesture recognition, the system turns the related text label into a speech to the deaf or mute user via a text-to-speech engine so that the gesture can be perceived by the user as an actual speech and therefore he/she can easily communicate with other people who do not know sign language.

2. Voice-to-Gesture Conversion

This is the reverse pipeline used to allow normal speakers to interact with the deaf or mute persons by translating voice commands into corresponding gesture outputs.

a. Microphone Input

The system starts by recording the words uttered by the user by a microphone. This is the raw output of the audio input, which the system has to read some meaning. Human speech comprises of tone variations, tempo and vocal depths, the microphone should be capable of capturing neat and consistent audio signals so that the system can effectively interpret the message. This will be necessary since quality audio enhances the accuracy of the speech recognition aspect that will come after it.

b. Speech Recognition

The speech recognition component analyzes the audio material that has been recorded by the microphone and recognizes the words that are spoken by the person with the help of trained linguistic and acoustic models. It divides the sound into meaningful patterns and compares them with the known speech data to identify the correct words. The step will make sure that the system captures the message of the user. Speech recognition should be reliable to develop an accurate and significant translation to gestures**.**

c. Sound-to-Text Conversion

At this point, the speech signal that has been processed is translated to a structured text format. Once the speech recognition module has extracted the spoken words, it filters the output produced by the system through correcting pronunciations, noise removal as well as mapping the identified speech to the nearest meaningful textual representation. This makes sure that the text version comes up to be what the user meant to say. Identification of a suitable gesture amongst the collection of gestures is then produced based on the resulting text, and the conversation between spoken language and sign-based communication is easily facilitated.

d. Call-Back Module

After the speech is translated into writing, the callback module will match this writing with the gesture in the dataset of the system. It is the decision making unit that chooses the right visual representation to the identified command. This mapping is necessary to put similarity between oral and sign output.

The module ensures that all the words identified are followed by an accurate and relevant gesture.

e. Gesture Display Output

This system then displays on the screen the chosen gesture to be interpreted by the deaf or mute user. This can be in the form of a fixed image, animation or a form of symbolism depending on the design. The read out is displayed on a clear screen and thus the message being spoken is easily understood upon translation. This last product is the final output that fulfills the communication process in converting speech to a form of accessible sign language.

3. Overall Purpose

The whole mechanism is made to eradicate communication barriers between the hearing and those with the speech or hearing impairments. The system offers a real-time translation interface between the spoken language and the sign language by incorporating computer vision, deep learning, speech recognition, and gesture rendering. This corresponds to the purpose of the abstract that allows the interaction of both groups using AI technologies in the form of a smooth, inclusive, and intuitive way.

12. CONCLUSION

In summary, the suggested Deaf and Dumb Communicator system represents a good form of technology to facilitate the communication barrier between the hearing- and speech- impaired and the rest of the population. Through the use of computer vision algorithms that process live video stream and the deep learning algorithms, namely the convolutional neural networks, to identify sign gestures, the system is able to translate sign language into a readable text and an audible audio file. This facilitates real time and substantive interaction between deaf or mute persons and the individuals who are not conversant with the sign language.

Facial emotion recognition increases the system further as non-verbal cues are captured, which are crucial in expressing intent and other emotional contexts leading to more natural and expressive communication. The multimodal system is a better recognizer and more usable than the single-modality system of the traditional systems. Also, the automated translation system helps to eliminate the need of human interpreters which facilitates easier communication that is also cheaper and can be scaled. In general, the work is beneficial in the direction of inclusive human-computer interaction since it illustrates the application of computer vision and deep learning to assistive technologies in the real world. The suggested system can enhance the social inclusion, education, and communication at the workplace of people with hearing impairment and speech disorders. As the system can be improved with additional support of more sign languages, real-life resilience, and mobile implementation, and can be used by more users to participate in a larger contributor to an inclusive and accessible digital society.

REFERENCES

1. Rojas, D. C., & Orozco, J. (2018). "A Review of Sign Language Recognition Systems." International Journal of Computer Applications, 182(16), 18-22. DOI: 10.5120/ijca2018916820
2. Yu, K., & Miao, Z. (2020). "Deep Learning for Gesture Recognition: A Review." IEEE Transactions on Human- Machine Systems, 50(5), 433-444. DOI: 10.1109/THMS.2020.2983459
3. Kafai, A., & Makarova, M. (2021). "Facial Emotion Recognition Using Deep Learning Techniques: A Review." Journal of King Saud University - Computer and Information Sciences. DOI: 10.1016/j.jksuci.2021.01.007
4. Ganaie, S. A., & Shah, S. A. (2019). "Recent Trends in Text-to-Speech Synthesis." Journal of King Saud University - Computer and Information Sciences, 31(1), 5-20. DOI: 10.1016/j.jksuci.2017.07.001
5. Alkhazaleh, R., & Ali, A. (2022). "RealTime Sign Language Recognition Using Deep Learning Techniques." Journal of Real-Time Image Processing, 19(2), 309-322. DOI: 10.1007/s11554-022-01209-4
6. D'Mello, S. K., & Graesser, A. C. (2015). "Multimodal Communication: Combining Speech, Gesture, and Emotion." Human- Computer Interaction, 30(4),346-
7. 370. DOI: 10.1080/07370024.2014.974087
8. Zha, Z., & Wu, Z. (2021). "A Survey on Sign Language Recognition: Challenges, Opportunities, and Future Directions." IEEE Transactions on Multimedia, 23, 2812-2831. DOI: 10.1109/TMM.2021.3071013
9. Poria, S., & Mihalcea, R. (2017). "Emotion Recognition in Video Using Multimodal Features." IEEE Transactions on Affective Computing, 8(1), 27-36. DOI: 10.1109/TAC.2016.2616353
10.]Hinton, G., Deng, L., Yu, D., et al. (2012). "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups." IEEE Signal Processing Magazine, 29(6), 82-97. DOI: 10.1109/MSP.2012.2205597
11. Mirończuk, M., & Pękala, M. (2021). "Gesture Recognition Systems: A Review of the State of the Art."
12. Sensors, 21(4), 1345. DOI: 10.3390/s21041345
13. Pu, J., Zhou, W., & Li, H. (2019). "Iterative Alignment Network for Continuous Sign Language Recognition." IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4165–4174. DOI: 10.1109/CVPR.2019.00429
14. Camgoz, N. C., Hadfield, S., Koller, O., & Bowden, R.(2017). "SubUNets: End-to-End Hand Shape and Continuous Sign Language Recognition." IEEE International Conference on Computer Vision (ICCV), 3075–3084. DOI: 10.1109/ICCV.2017.333
15. Huang, J., Zhou, S. K., & Metaxas, D. (2011). "A Review of Hand Gesture Recognition Based on Computer Vision Techniques." IEEE International Conference on Multimedia and Expo, 1–6. DOI: 10.1109/ICME.2011.6012171
16. Molchanov, P., Gupta, S., Kim, K., & Kautz, J. (2016). "Hand Gesture Recognition with 3D Convolutional Neural Networks." IEEE Conference on Computer Vision and Pattern Recognition Workshops, 1–7. DOI: 10.1109/CVPRW.2016.201
17. Zhou, Z., & Chellappa, R. (2006). "From Hand Gestures to Human Activity: Recognition Using Shape and Motion Features." IEEE Transactions on Image Processing,15(6), 1833–

1846. DOI: 10.1109/TIP.2006.873444
18. Simonyan, K., & Zisserman, A. (2014). "Two-Stream Convolutional Networks for Action Recognition in Videos." Advances in Neural Information Processing Systems (NeurIPS), 568–576.
19. Sutskever, I., Vinyals, O., & Le, Q. V. (2014). "Sequence to Sequence Learning with Neural Networks." Advances in Neural Information Processing Systems, 3104–3112.
20. Schuster, M., & Paliwal, K. K. (1997). "Bidirectional Recurrent Neural Networks." IEEE Transactions on Signal Processing, 45(11), 2673–2681. DOI: 10.1109/78.650093
21. Koller, O., Forster, J., & Ney, H. (2015). "Continuous Sign Language Recognition: Towards Large Vocabulary Statistical Recognition Systems Handling Multiple Signers." Computer Vision and Image Understanding, 141, 108–125. DOI: 10.1016/j.cviu.2015.08.004
22. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. (Widely cited foundational reference for CNNs, RNNs, and DL theory)