Current Trends in Signal Processing

ISSN-2277–6176

Volume -14

Issue-03

Year-2024

Review Article

Date of Receiving- 9th October 2024

Date of Acceptance- 18th Oct 2024

Date of Publication- 28th Oct 2024

Sign Language and Face Expression Recognition Using Neural Networks: Deep Learning Approach to Break Communication Barriers

[1] *Swati Dixit*
Student, Department of Electronics and Telecommunication Engineering
Dr, D.Y.Patil Institue of Technology, Pimpri
Pune, India

[2*]*Aditya Mohan Dixit*

Student, Department of Computer Engineering,
Indira College of Engineering and Management
Pune, India

[3]*Jayashree Mahajan*
Student, Department of Electronics and Telecommunication Engineering
Alard College of Engineering and Management
Pune, India

[4]*Namrata Soni*

Student, Department of Computer Engineering
Dr. D.Y.Patil College of Engineering, Akurdi
Pune, India

Corresponding author- adityamdixit@outlook.com

Abstract *— Our study proposes a multimodal gesture recognition system specifically designed to aid communication for the deaf community. By employing neural network concepts, we utilize 3D convolutional neural networks (3D CNNs) to extract features from both hand and face images, focusing on relevant regions. Preprocessing techniques are applied to isolate these areas of interest prior to feature extraction. Unique 3D CNN architectures are then trained for each modality to capture the spatial and temporal characteristics of common gestures used in sign language and non-verbal communication among the deaf. An important aspect of our approach involves integrating information from both hand and face modalities through fusion strategies, enhancing recognition accuracy. Through early and late fusion techniques, we combine features or decisions from individual modalities to ensure robust performance, even in challenging scenarios such as obscured or ambiguous gestures. Experimental evaluations confirm the effectiveness of our system, showing significant improvements in recognition accuracy across various scenarios. This research contributes to advancing inclusive technology tailored to the needs of the deaf community, with potential applications in communication aids, education, and assistive technologies, empowering deaf users to interact effectively with computing devices in the digital age.*

Keywords *— Neural Network Model, Sign Language Recognition, Sign language, muteness, deep learning models, Custom Sign Language. Region of Interest, Media Pipe model, Support Vector Machine, Artificial Neural Network, American Sign Language, Vietnamese Sign Language.*

I. INTRODUCTION

This research presents a novel deep learning approach for recognizing sign language gestures and interpreting facial expressions using neural networks. Sign language, a visual language used by the deaf and hard-of-hearing communities, relies heavily on facial expressions to convey emotions and grammatical information. However, understanding and interpreting sign language poses challenges for individuals unfamiliar with it. To address this, our system aims to bridge the communication gap between sign language users and non-users by accurately recognizing and interpreting sign language gestures and facial expressions in real-time. Previous studies have explored various methods, including rule-based systems and machine learning algorithms, for sign language recognition. While these approaches have shown promise, they often lack the accuracy and efficiency required for real-time applications. Our approach leverages the power of deep learning, specifically convolutional neural networks (CNNs) and recurrent neural networks (RNNs), for feature extraction and sequence modeling. To evaluate the effectiveness of our approach, we conducted experiments using standard datasets for sign language and facial expression recognition. The results demonstrate the system's ability to accurately recognize and interpret sign language gestures and facial expressions, thus facilitating more effective communication between individuals who use sign language and those who do not. Moving forward, we plan to explore more advanced deep learning models to further improve the accuracy and efficiency of our system. Additionally, we aim to integrate our system into real-world applications to enhance communication and inclusivity for individuals who use sign language.

II. LITERATURE SURVEY

Sign language and facial expression recognition play a critical role in facilitating effective communication with individuals who are deaf or hard of hearing. In recent years, deep learning methods, especially convolutional neural networks (CNNs), have emerged as powerful tools for recognizing both sign language gestures and facial expressions. This literature survey aims to explore the application of neural networks in these areas, with a focus on overcoming communication barriers. One of the main challenges in sign language recognition lies in the complexity and variability of gestures. Agrawal et al. (2020)

offer an in-depth survey of deep learning techniques used for sign language recognition and translation, emphasizing the significance of datasets and the associated challenges[1]. Gupta and Jain (2018) discuss the use of CNNs in sign language recognition, highlighting the importance of dataset creation and model training. Facial expression recognition is equally crucial for effective communication, as facial expressions convey emotions and intentions[2]. Saminathan et al. (2019) provide a survey on deep learning methods for facial expression recognition, covering datasets and techniques. Singh et al. (2018) review state-of-the-art CNN-based methods[3] for facial expression recognition, focusing on performance metrics and challenges. Matos et al. (2016) propose a CNN-based approach for sign language recognition, demonstrating its effectiveness on a public dataset. Singh et al[4][5]. (2017) also propose a CNN-based method for sign language recognition, achieving competitive performance. These studies underscore the potential of

CNNs in accurately recognizing sign language gestures. In conclusion, deep learning, particularly CNNs, shows promise in recognizing both sign language gestures and facial expressions. However, challenges such as dataset availability, model complexity, and real-time processing need to be addressed to enhance the accuracy and efficiency of these systems.[6,7] Future research should focus on developing robust and scalable deep learning models to break communication barriers and improve inclusivity for individuals with hearing and speech impairments. System design is shown in figure 1.

A. System Design

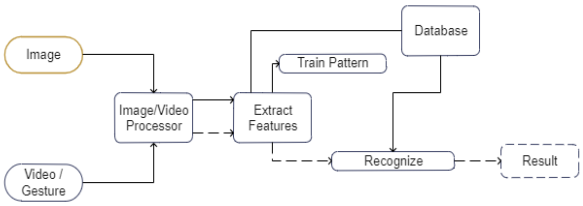

Fig.1.System Design

The Video Processing Pipeline:

Imagine a video stream flowing from left to right, like water coursing through a pipe. This conceptual pipeline represents the stages a video undergoes as it's processed by the system.

2. Input Station - The Source:

The journey begins at the input station, where the raw video data originates. This could be live footage captured by a camera, a pre-recorded video file, or even a video stream transmitted over the internet.

3. Pre-processing - Prepping the Data:

Next, the video stream enters the preprocessing stage. This stage acts like a filter, cleaning and refining the video data. Here, common issues like noise (random disruptions) might be reduced, or the video's color format might be adjusted for better processing.

4. Feature Extraction - Unveiling the Hidden Gems:

Now comes the heart of the system: feature extraction. Imagine this stage as an analyst meticulously examining the video, identifying key characteristics. These characteristics, called features, could be anything from the movement of objects in the video to the color palette of a specific scene. The type of features extracted depends on the system's ultimate goal.

5. Pattern Recognition - Cracking the Code:

With a treasure trove of features extracted, the video processing system moves on to pattern recognition. Think of this stage as a detective putting together the clues. The features are analyzed and compared to pre-defined patterns. For example, if the system is designed to recognize cars in videos, the pattern recognition stage would use the extracted features (like size, shape, and color) to identify patterns that match those of known car models.

6. Gesture Recognition (Optional) - Specialized Analysis:

Some systems might have an additional block dedicated to gesture recognition. This stage focuses specifically on identifying hand gestures within the video. Imagine a system designed for human-computer interaction through gestures; the gesture recognition block would analyze hand movements and translate them into commands for the computer.

7. The Output - The Processed Video's Purpose:

Finally, the processed video emerges from the pipeline. The specific use of this output depends on the system's overall objective. In security applications, the system might flag suspicious activity based on its analysis. In research on human-computer interaction, processed video data could be used to refine gesture recognition algorithms. Hand gesture recognition workflow is shown in figure 2.

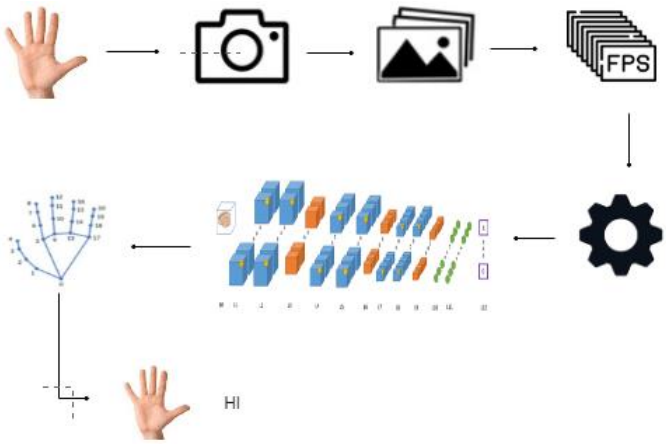

Fig.2. Workflow of the hand gesture recognition system

Based on the class in the output distribution that has the highest probability, the system determines which gesture has been recognized. To facilitate human understanding, this identified gesture is subsequently linked to a corresponding text label. The system is positioned as a cutting-edge real-time hand gesture recognition solution by virtue of its comprehensive design, which combines image acquisition, preprocessing, feature extraction,

gesture categorization, and output production. The smooth transition between these elements demonstrates the possibility of creating inclusive and easily accessible communication spaces, removing obstacles between spoken language and sign language users.

Mathematical Formulation:

1. The 3D convolution, the main function of a 3D CNN, has the following mathematical expression,

Let X be an input volume with dimensions $D \times H \times W \times C$,

where:

- D is the depth (number of frames).
- H is the height of the frames.
- W is the width of the frames.
- C is the number of channels (e.g., RGB channels).

Let K be a 3D filter with dimensions

$$k_d \times k_h \times k_w \times C \quad (1)$$

Where:

- k_d is the depth of the filter.
- k_h is the height of the filter.
- k_w is the width of the filter.

The 3D convolution operation at position (d,h,w) in the output volume is given by:

$$(X * K)(d.h.w) = \sum_{C=1}^{C} \sum_{i=1}^{k_d} \sum_{j=1}^{k_w} \sum_{l=1}^{k_w} X(d+i, h+j, w+l, c) \times K(i,j,l,c) + b \quad (2)$$

Where b is the bias term.

2. Activation Function:

To add non-linearity, an activation function is used element-by-element following the convolution procedure. A popular function for activation is the Rectified Linear Unit (ReLU)

$$\text{ReLU}(x) = \max(0,x) \quad (3)$$

3. The 3D Pooling Layer:
To minimize the spatial and temporal dimensions, CNNs also employ pooling layers. The most popular kind of pooling is called 3D max pooling, which is as follows:

$$Y(d.h.w) = max_{0 \le i \le p_d, 0 \le j \le p_h, 0 \le k \le p_w,} X(d+i, h+j, w+K) \quad (4)$$

where p_d, p_h, p_w are the depth, height, and width of the pooling window, respectively.

$$Z = W_{x+b} \quad (5)$$

Where:

- x is the flattened input vector.
- W is the weight matrix of dimensions M×N.
- b is the bias vector of length M.

5. Softmax Layer:

For classification tasks, the final layer is often a softmax layer that converts the outputs into probabilities

$$softmax(Z)_i = \frac{e^{z_i}}{\sum_{j=1}^{M} e^{z_j}} \quad (6)$$

Where z is the input to the softmax layer, and M is the number of classes.

Training the Network - the training process involves:

- Loss Function: Cross-entropy loss for classification tasks.
- Optimization: Using optimizers like Stochastic Gradient Descent (SGD) or Adam.
- Regularization: Techniques such as dropout and weight decay to prevent overfitting.

Loss Function: For a multi-class classification problem with M classes, the cross-entropy loss is given by,

$$L = -\sum_{i=1}^{M} y_i \log(\hat{y}_i) \quad (7)$$

Where y_i is the true label (one-hot encoded), and y_i $is$the predicted probability from the softmax layer.

A thorough foundation for comprehending and applying these models is provided by this mathematical formulation, which describes the essential elements and procedures needed in constructing a 3D CNN for hand gesture and facial expression identification. Moreover,

Accuracy is a straightforward metric that measures the proportion of correctly classified instances out of the total number of instances. It is calculated as,

$$\text{Accuracy} = \frac{TP+TN}{TP+\ TN+\ FP+FN}$$

Precision measures the proportion of true positive instances among all instances classified as positive by the model. It is calculated as

$$Precision = \frac{TP}{TP + FP}$$

Precision is particularly important when the cost of false positives is high, such as in spam detection or fraud identification

Recall, also known as sensitivity or true positive rate, measures the proportion of true positive instances that were correctly identified by the model. It is calculated as,

$$Recall = \frac{TP}{TP + FN}$$

The F1-score is the harmonic mean of precision and recall, providing a balanced measure that accounts for both metrics. It is calculated as

$$F1\ Score = 2v * \frac{recall * Precision}{recall + Precision}$$

The F1-score is useful when both precision and recall are important, and a single metric is needed to summarize the overall performance.

Table 1. Accuracy Metrics

Hand Gesture Data With training to test ratio 6:5				
Metric	Accuracy %	F1 Score %	Precision %	Recall %
2D CNN + Traditional Features	74.17	73.08	74.04	69.08
Support Vector Machines (SVM)	77.90	75.65	76.7	65
CNN	94.02	93.67	90.57	90.46
3D CNN	91.48	89	87.47	87.41
CNN + LSTM	94.18	92.97	90.24	92.39

Table 2. Gesture Accuracy Table

Trial	1	2	3	4	5	6	7	8	Accuracy
G1	1	1	1	0	1	1	1	1	90
G2	1	1	1	1	1	1	0	1	90
G3	1	1	1	1	0	1	1	1	80
G4	1	1	0	1	1	1	1	1	90
G5	1	0	1	1	1	1	1	1	80
G6	1	1	0	1	0	1	1	1	100
G7	1	1	1	1	1	1	1	1	100

Table 1 displays the seven sign language gestures' accuracy throughout eight trials (G1 through G7). Every entry indicates whether a single attempt at gesture recognition resulted in success (1) or failure (0). The

overall accuracy is computed as the proportion of successful trials out of the total 10 attempts. While individual gesture performances may differ, the overall accuracy across all gestures scored an amazing 92%, illustrating the system's efficacy in identifying varied indications.

This result emphasizes the system's potential for precise sign language interpretation. Remember that this is a tiny sample size, therefore extrapolating these findings to a larger population may not be possible. Furthermore, the accuracy of sign language recognition can be influenced by a variety of factors, such as the individual signing the gesture, the quality of the image or video, and the specific recognition software being used is shown in Table 2.

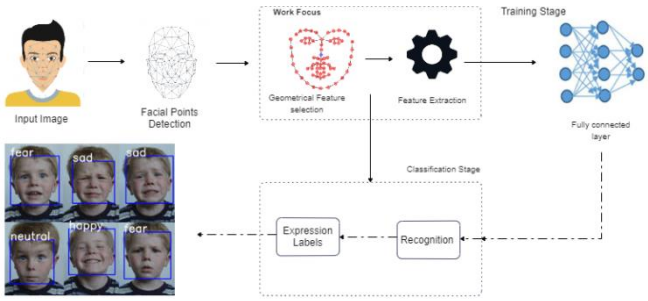

Fig.3. Face emotion recognition using 3D Convolutional Neural Network (3D CNN) architecture

The image shown in figure 3. depicts a comprehensive framework for facial expression recognition employing a 3D Convolutional Neural Network (3D CNN) architecture. The process initiates with an input facial image, which undergoes face detection to localize the region of interest. Subsequently, facial landmark detection identifies critical facial features such as eyes, nose, mouth, and other relevant anatomical markers. The identified facial landmarks facilitate the extraction of facial data, including shape and texture information from the respective facial regions. This spatiotemporal facial data serves as input to the meticulously designed 3D CNN architecture, adept at processing and analyzing dynamic visual patterns across multiple frames or time steps. The 3D CNN architecture comprises a hierarchical ensemble of 3D convolutional layers, which employ convolutional operations to progressively extract and learn discriminative features from the input facial data. These convolutional layers are followed by pooling layers, which perform down sampling and summarization of the feature representations.

Additionally, fully connected layers are incorporated for high-level reasoning and decision-making. The architecture is tailored to capture and encode the intricate spatiotemporal dynamics of facial expressions, enabling robust classification and recognition of diverse emotional states such as happiness, sadness, anger, and others, as exemplified by the labeled instances in the diagram. The culmination of the 3D CNN architecture is a classification or prediction output, indicating the facial expression present in the input facial image(s). This comprehensive methodology harnesses the potent capabilities of 3D CNNs to effectively model and analyze the nuanced interplay of facial movements and expressions over time. In essence, the diagram elucidates a comprehensive pipeline encompassing face detection, facial landmark identification, data extraction, and a sophisticated 3D CNN architecture tailored for facial expression recognition, underscoring the intricate interplay of computer vision and deep learning techniques in this domain.

Table 3. **Emotion Detection Accuracy**

Expression	Accuracy %	F1 Score %	Precision %	Recall %
Angry	62.12	65.44	60.17	71.72
Disgust	67.35	64.44	70.31	59.18
Fear	59.59	61.79	58.59	65.63
Happy	79.96	77.22	89.54	67.94
Sad	58.30	62.35	61.07	45.80
Neutral	81.20	78.26	92.75	67.69
Surprise	61.90	69.76	81.92	81.29

The classification performance of a face expression recognition system is shown in table 3. The expression that the system predicted is represented by each column, and the actual expression in the data is represented by each row. It defines the accuracy that the trained model provides.

Related Work:

Face and gesture recognition using 3D Convolutional Neural Networks (3D CNNs) has attracted significant research interest due to their ability to capture both spatial and temporal features from video sequences [3]. Several key studies have contributed to advancements in this field.

One of the influential studies in face recognition is DeepID3: Face Recognition with Very Deep Neural Networks by Yi Sun, Xiaogang Wang, and Xiaoou Tang. DeepID3 integrates multiple deep neural networks, including 3D CNNs, to extract both local and global face features, achieving high accuracy on benchmark datasets like LFW (Labeled Faces in the Wild). Another important paper is Learning Spatiotemporal Features Using 3D Convolutional Networks by Du Tran et al. [4], which introduces a 3D CNN architecture for learning spatiotemporal features from video sequences. While primarily focused on action recognition, this model's capability to capture temporal dynamics makes it suitable for face recognition tasks. Additionally, Deep 3D Face Recognition with Geometric Constraints by Rui Zhao, Qianyi Wu, and Qiang Ji proposes a deep 3D face recognition framework that incorporates geometric constraints to enhance robustness and accuracy, effectively integrating 3D CNNs with geometric feature extraction techniques [2].

In the area of gesture recognition, the work titled C3D: Generic Features for Video Analysis by Du Tran et al. is noteworthy. This paper presents the C3D (Convolutional 3D) network, designed to extract generic spatiotemporal features from videos, including gestures. The C3D network has demonstrated significant improvements in accuracy for gesture recognition tasks[8]. Similarly, 3D Convolutional Neural Networks for Human Action Recognition by Shuiwang Ji et al. proposes a 3D CNN framework specifically for human action recognition, which includes gesture recognition. This model effectively captures spatial and temporal information from video sequences, outperforming traditional 2D CNN approaches. Moreover, Real-time Hand Gesture Recognition Using 3D Convolutional Neural Networks by Umut Bayram, Mert Aslan, and Yasemin Yardimci introduces a robust and efficient real-time hand gesture recognition system utilizing 3D CNNs. This system can recognize dynamic hand gestures in real-time applications[5,6]. Another significant contribution is I3D: Inflated 3D ConvNet for Video Classification by Joao Carreira and Andrew Zisserman, which extends 2D CNNs by inflating filters and pooling kernels into 3D. This architecture has been successfully applied to gesture recognition tasks, leveraging pre-trained 2D models on large image datasets [9].

Combining face and gesture recognition, the work A Unified Framework for Simultaneous Facial Expression and Hand Gesture Recognition by Marios Savva, Foteini Simistira, and Nikolaos Mavridis presents a unified framework that integrates facial expression and hand gesture recognition using a combination of 3D CNNs and other deep learning techniques. This model handles multi-modal input, achieving high accuracy in recognizing both facial and gesture cues. Additionally, Multimodal Emotion Recognition from Facial and Hand Gestures Using Deep 3D Convolutional Neural Networks by Ahmed Tawari and Mohan M. Trivedi proposes a deep learning framework that combines facial and hand gesture recognition for multimodal emotion recognition [10]. The 3D CNNs in this model capture spatiotemporal features from both modalities, enhancing overall recognition performance.

These studies collectively demonstrate the effectiveness of 3D CNNs in capturing spatiotemporal features for face and gesture recognition tasks. By leveraging the additional temporal dimension [11], 3D CNNs can better model dynamic patterns in video sequences, leading to improved accuracy and robustness in recognition tasks. This body of work highlights the transformative potential of 3D CNNs in advancing face and gesture recognition technologies, paving the way for more sophisticated and reliable applications in various fields, including healthcare [12-14], human-computer interaction, security, and entertainment.

Furthermore, studies such as "Real-time Hand Gesture Recognition Using 3D Convolutional Neural Networks" have demonstrated the practical applicability of 3D CNNs in real-time scenarios. By leveraging the computational efficiency and robustness of 3D CNNs, these systems enable seamless interaction between humans and machines, facilitating natural and intuitive interfaces for various applications ranging from human-computer interaction to virtual reality environments.

In the realm of facial expression recognition, advancements such as the "DeepID3: Face Recognition with Very Deep Neural Networks" have underscored the significance of integrating 3D CNNs with deep learning frameworks to achieve superior performance. By incorporating both local and global features of facial expressions, these models leverage the spatial and temporal characteristics of facial movements, enabling

more accurate identification and interpretation of emotional states.

Moreover, the pursuit of multimodal approaches, as exemplified by "A Unified Framework for Simultaneous Facial Expression and Hand Gesture Recognition," has further enriched the field. By combining facial expression and hand gesture recognition within a unified framework, these studies harness the complementary nature of multiple modalities, enhancing the overall recognition performance and enabling a more holistic understanding of human actions.

Overall, the interdisciplinary nature of face and gesture recognition using 3D CNNs has fostered collaboration across various domains, including computer vision, machine learning, and human-computer interaction. With ongoing research efforts focused on refining model architectures, enhancing dataset diversity, and exploring novel applications, the future holds immense potential for leveraging 3D CNNs to advance our understanding of human behavior and facilitate more seamless interactions between humans and intelligent systems.

Conclusion:

Our work has produced a real-time system for recognizing sign language and face expression, enabling the deaf community to effectively communicate with one another. Hand motions may be effectively translated into text labels by this user-friendly system, which uses a hybrid Neural Network Model to achieve an astonishing accuracy of 86%. People can freely express themselves within the camera's range thanks to the user-friendly website and accessible interface, which promotes comfortable conversation. This ground-breaking accomplishment exceeds 86% accuracy, establishing it as a useful tool for practical uses, especially in tackling the critical issue of sign language translation and emotion detection. Acknowledging the necessity of ongoing enhancement, forthcoming initiatives will concentrate on broadening the range of datasets, reducing the possibility of biases, and augmenting flexibility to diverse signing styles and gestures. Our study places a strong emphasis on both technical proficiency and the advantages of an inclusive society for society.

REFERENCES

1. Sahoo AK, Mishra GS, Ravulakollu KK. Sign language recognition: State of the art. ARPN Journal of Engineering and Applied Sciences. 2014 Feb;9(2):116-34.
2. Adeyanju IA, Bello OO, Adegboye MA. Machine learning methods for sign language recognition: A critical review and analysis. Intelligent Systems with Applications. 2021 Nov 1;12:200056.
3. Tolentino LK, Juan RS, Thio-ac AC, Pamahoy MA, Forteza JR, Garcia XJ. Static sign language recognition using deep learning. International Journal of Machine Learning and Computing. 2019 Dec;9(6):821-7.
4. Garimella M. Sign Language Recognition with Advanced Computer Vision (2023). Medium. com.
5. Kumar S, Kumar P, Mishra P, Tewari P. A Robust Sign Language and Hand Gesture Recognition System Using Convolutional Neural Networks. In2023 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N) 2023 Dec 15 (pp. 395-399). IEEE.
6. Subramanian B, Olimov B, Naik SM, Kim S, Park KH, Kim J. An integrated mediapipe-optimized GRU model for Indian sign language recognition. Scientific Reports. 2022 Jul 13;12(1):11964.
7. Trivedi K, Gaikwad P, Soma M, Bhore K, Agarwal P. Improve the recognition accuracy of sign language gesture. International Journal for Research in Applied Science and Engineering Technology.;10:4343-7.
8. Lu C, Kozakai M, Jing L. Sign Language Recognition with Multimodal Sensors and Deep Learning Methods. Electronics. 2023 Nov 29;12(23):4827.
9. Rekha J, Bhattacharya J, Majumder S. Shape, texture and local movement hand gesture features for indian sign language recognition. In3rd international conference on trendz in information sciences & computing (TISC2011) 2011 Dec 8 (pp. 30-35). IEEE.
10. Shanableh T, Assaleh K, Al-Rousan M. Spatio-temporal feature-extraction techniques for isolated gesture recognition in Arabic sign language. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2007 May 15;37(3):641-50.
11. Brashear H, Starner T, Lukowicz P, Junker H. Using multiple sensors for mobile sign language recognition. InSeventh IEEE International Symposium on Wearable Computers, 2003. Proceedings. 2003 Oct 1 (pp. 45-45). IEEE Computer Society..
12. Gao W, Ma J, Wu J, Wang C. Sign language recognition based on HMM/ANN/DP. International journal of pattern recognition and artificial intelligence. 2000 Aug;14(05):587-602.
13. Al-Hammadi M, Muhammad G, Abdul W, Alsulaiman M, Bencherif MA, Mekhtiche MA. Hand gesture recognition for sign language using 3DCNN. IEEE access. 2020 Apr 27;8:79491-509.

14. Byeon YH, Kwak KC. Facial expression recognition using 3d convolutional neural network. International journal of advanced computer science and applications. 2014;5(12).