RESEARCH PAPER

Gesture Recognition and Voice Control System

Ms. Latika Sharma[1], Shivam Shukla[2], Rohan Mittal[3], Vaibhav[4], Varun Sharma[5]
Computer Science and Information Technology, KIET Group of Institutions, Ghaziabad, India
[1]Email: latika86@gmail.com
[2]Email: shivamshuklah@gmail.com
[3]Email: rohanmittalsdps@gmail.com
[4]Email: vaibhavchahar7@gmail.com
[5]Email: varun225176@gmail.com

ABSTRACT
This is a real-time Gesture Recognition and Sound Control System that is aimed at allowing a personal computer to be controlled by hand gestures without touching the computer. The system records live video feeds by use of web camera and tracks the hand movement and position of fingers by use of computer vision algorithms. The system recognises specific predefined gestures and maps them to particular computer commands e.g. volume control, media playback, cursor movement, application switching, and system navigation by extracting key hand landmarks and analysing the pattern of gestures with deep learning models which are trained on a wide range of diverse datasets that are captured under different backgrounds and different lighting conditions. The accuracy and low latency of the system are proven by the experimental results and make it possible to implement the system in practice. The method is more accessible, convenient to the user, and offers an easy interface to interact with the computers, especially to users with physical needs.

Index Terms: Gesture Recognition, Human-Computer Interaction, MediaPipe, Deep Learning, Hand Tracking, PC Automation, Real-Time System.
How to cite this article: Sharma L, Shukla S, Mittal R, Vaibhav, Sharma V. Gesture Recognition and Voice Control System. Int J Drug Deliv Technol. 2026;16(52s): 754-760. DOI: 10.25258/ijddt.16.52s.92
Source of support: Nil.
Conflict of interest: None.

I. Introduction

Gesture recognition involves the use of cameras and the processing of images to detect the movement[1],[3] of hands. Voice recognition systems change speech into writing and perform actions. Together, these systems offer an effective multimodal interface which is user-friendly and available.The present project deals with the creation of a gesture and voice-based control system. The reason to choose this work is to facilitate a touchless interaction, particularly with physically challenged users and enhance user convenience in daily use applications like smart home and automation systems.Though there have been previous studies on scream detec- tion, there are still a number of challenges with the current methods. Conventional frameworks are vulnerable to back- ground noise, whereas other more recent deep-learning frame- works can only analyse a smaller number of features of the signal, such as local spectral features or long-range temporal structure. In practice, there are a wide range of screams based on duration, intensity, and conditions of recordings and it is not always easy to detect them.

Problem Statement: Conventional sound control systems use physical evidence like remote controls, buttons, or touch panels. The issue is to come up with a system that would enable users to manipulate sound features without touching any hardware device and with the help of hand gestures only.

Contributions: The main contributions of this work are as follows:

- Designed and developed a real-time hand gesture recog- nition system for controlling sound functions without physical contact.
- Implemented an efficient image acquisition and prepro-cessing pipeline to enhance gesture detection accuracy.
- Integrated the recognized gestures with audio con-trol operations, enabling intuitive and touchless hu- man–computer interaction.

III. Related Work

The field of gesture recognition and sound control systems is not a recent one as their significance in terms of human-computer interaction and intelligent interface design has been on the rise over several years. The initial studies in this field were largely based on the hardware solutions, including data

II. Literature Review

TABLE I

Comparative Analysis of Gesture Recognition Systems

S. r. No	Authors	Title	Key Findings
1	Neel Kamal et al. (2019)	[1]Indian Sign Language Gesture Recognition using Image Processing and Deep Learning	The system achieved 98.81% accuracy for Indian Sign Language gestures and 97.71% for American Sign Language, promising enhanced communication for the speech impaired.
2	Natrajan et al. (2022)	[12]Development of an End-to-End Deep Learning Framework for Sign Language Recognition, Translation, and Video Generation	This paper presents innovative solutions for real-time Sign Language recognition, translation, and video generation, achieving over 95% classification accuracy and significant improvements in recognition precision and visual fidelity, as evidenced by various evaluation metrics.
3	T. M Reddy et al. (2023)	Sign Language Recognition Using OpenCV and Convolutional Neural Networks	This study demonstrates CNN's efficacy in recognizing diverse sign language gestures, enabling rapid and precise communication without translation needs, thus enhancing accessibility for the speech and hearing impaired globally.
4	Nun Tadh	A Real-Time American Sign [1]Language Recognition System	This paper introduces a pioneering real-time ASL recognition system achieving remarkable performance using deep learning techniques and real-time processing strategies.
5	G. A Rao et al. (2018)	Deep convolutional neural networks for sign language recognition	[21]This paper presents a CNN-based approach for Indian sign language recognition, leveraging self-operated SLR mobile application data and a newly created dataset. Achieving a recognition rate of 92.88%, it signifies a significant stride in sign language recognition technology.
6	Amrita Thakur et al. (2020)	[5]Real time sign language recognition and speech generation	This paper underscores the significance of Sign Language Recognition systems, employing neural networks for accuracy and user-friendliness, enabling independent and seamless two-way communication for the hearing impaired.

gloves[14],[20], wearable sensors, and infrared markers to record the movements of hands. These systems cost a lot, were uncomfortable, and unfeasible to use in ordinary life, which would not give the right gesture data. The necessity to develop more natural and affordable solutions is what prompted the researchers to consider cameras-based vision-based systems of gesture recognition that considerably decreased hardware-dependency and provided better flexibility and comfort to the user.

Gesture recognition has been an important area of research that was revolutionized by the evolution of deep learning. Neural networks Convolutional neural networks (CNNs) also allowed automatic feature recognition on raw images, without manually designing features. A number of authors utilized CNN-based models to detect hand gestures with greater accuracy[5],[16] and better generalization. These methods had good performance, especially when used with controlled datasets, but usually needed numerous training examples and extensive computations.

Recent work has had a more emphasis on real-time performance and usability, with lightweight models and optimized pipelines that can run on consumer-grade hardware[2],[4]. There were those that used depth cameras and stereo vision to enhance the accuracy of hand segmentation and gesture recognition and others that examined a hybrid of vision-audio solutions. Much as the depth-based systems were more precise, they were also expensive and more difficult to get so this hindered their general use. Therefore, RGB camera-based solutions operated on efficient algorithms were still being refined by many researchers to balance between accuracy and computational efficiency.

Fig. 1. Cursor movement

IV. Methodology

The system takes in real time video input in a camera and does some pre processing to clear off noise and improve image quality. The video frames are identified to extract the hand region and features considered to depict various gestures. To ensure that specific hand gestures are properly identified, these features are grouped with the help of a trained machine learning model. All identified gestures are associated with a particular sound control command, i.e., volume control or mute. The audio corresponding action is performed immediately and allows to control sound in real time and without touching anything.

A. Image Acquisition

The former involves the live video capture through a webcam. The video stream is broken down into single frames to

Fig. 2. Brightness Control

process it. An appropriate frame rate is applied in order to have a real-time smooth performance.

B. Image Preprocessing

Gesture recognition has been enhanced by preprocessing. The operations that are most frequently performed are:

Conversion from RGB to grayscale
Noise removal using Gaussian or median filters
Background subtraction
Image normalization and resizing

C. Hand Detection

The detection of hands is done by detecting the region of the hand in the image. Contour detection is common of which the largest contour is assumed to be the hand. The area of interest is cut out and analyzed further.

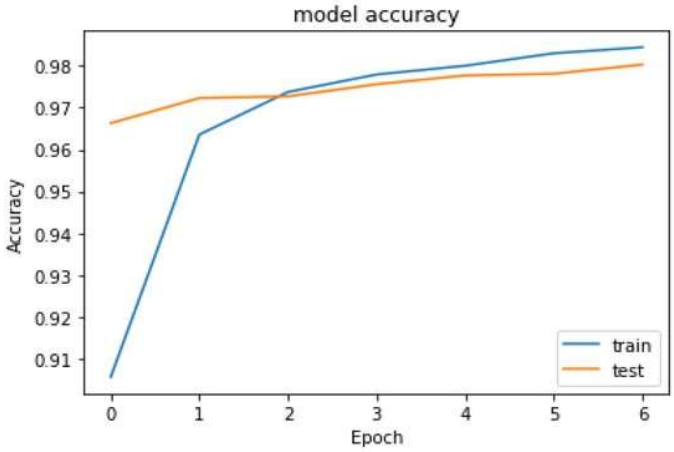

Fig. 3. This is my image caption

D. Feature Extraction

Important characteristics like the shape of the hands, area of the contours, and defects in the shapes of the convexities and the number of fingers are extracted. The aspects contribute to the distinction of various gestures.

E. Gesture Classification

Gesture classification is performed with the help of machine learning algorithms (K-Nearest Neighbors (KNN), Support Vector Machine (SVM), or Convolutional Neural Networks (CNN)[15],[16]). Gesture labeled data are used to train the model.

F. Sound Control Execution

After a gesture is detected, it is connected to a sound control operation. For example:

Open palm: Volume up
Closed fist: Volume down
Two fingers: Play or pause
Thumb up: Mute

V. Dataset Description

The data to be applied in this gesture recognition and sound control project is the images of hand gestures and video frames which are captured by a conventional RGB camera. It has several predetermined hand signs that relate to sound control commands like the volume up, volume down, mute, play and pause. The data is gathered among various users so as to be varied in terms of hand size, shape, orientation, and style of gesture execution. The photographs are taken in different lighting conditions and backgrounds to enhance the strength and the generalization property of the model. The samples of gestures are also labeled correctly in order to facilitate supervised learning in the training. The dataset is separated into training and testing sets, so the efficiency of gesture recognition and real-time functionality of the system could be assessed successfully.

VI. Experimental Setup

A. Implementation Details

The Gesture Recognition and Voice Control System is implemented under a modular architecture of software application to guarantee the flexibility, scalability, and ease of maintenance. Implementing the system is mainly in Python because it has a wide library support on computer vision as well as speech processing and machine learning. The implementation is centered on real-time performance, accuracy and compatibility with low cost hardware.

B. Hardware Setup

The system has very few hardware requirements. Visual input is captured through a standard webcam and audio input is captured through a microphone to allow gesture and voice commands respectively. Laptops and desktop systems are also often compatible with such devices, which is why the solution is cost-effective and can be deployed with ease. There is no need in special sensors or wearable gadgets.

C. Software Environment

The implementation of the software is done on general purpose operating system like windows or Linux. Image and video processing is done in libraries like OpenCV whereas audio input processing is done in speech recognition libraries. The machine learning models are deployed on popular frameworks that facilitate effective model training and inference.

D. Gesture Recognition Implementation

The technique involves gesture recognition that is applied through the use of computer vision. The webcam also captures video frames which are processed in real time. Every frame passes through preprocessing procedures like resizing, filtering noise, and filtering background. The detection and segmentation of the hands is done to isolate the area of interest. The segmented hand region is transformed into the relevant features like the hand shape, finger positions, and patterns of motion. The features extracted are then given to a trained classifier that recognizes the gesture performed. There is both a support of static and dynamic gestures. The gestures that are considered as static include gestures that are considered to be based on the position of hands but the gestures that are considered to be dynamic involve observation of motion over various frames. The system has been made such that it can deal with the changes in gesture performance and lighting.

E. Voice Control Implementation

The voice control is done with the help of speech recognition pipeline. The microphone is constantly monitored. The audio signal when speech is detected is preprocessed to eliminate noise and silence and reduce the signal. The extraction of features is consequently done in order to identify pertinent speech features. A speech-to-text engine is used to convert the processed audio into a text. Known text is compared to set of commands. The system facilitates easy and well-defined voice commands to allow the system to achieve high recognition. Adding new commands to the design is made easily as they are needed.

F. Command Mapping and Execution

A command mapping module is used to map the known gestures and voice commands to system actions. This module also makes it possible to ensure that every known input activates the right functioning. Where gesture and voice input are received at the same time, priority rules or contextual logic are used to deal with conflict situations. When a command is confirmed, the control module performs the respective action, which can be to control a device, navigate an interface or even activate a system action. Visual representations or sounds are used to give feedback to the user to confirm that he or she has completed a task successfully.

G. Performance and Testing

To test the system, the system is subjected to varying lighting and noise conditions to test the accuracy and robustness of the system. Optimization measures like frame rate control and effective feature extraction are employed to make sure that it operates in real-time. The modular architecture enables one to test and enhance individual components.

VII. Results and Discussion

Gesture Recognition and Voice Control System was also applied and tested to measure its performance, accuracy, reliability, and real time responsiveness. The findings indicate that the suggested system was successful in facilitating the natural and non-contact interaction between the user and the machine through hand gestures and voice commands. The system was also put to test in different conditions to determine its strength and its applicability in real life.

A. Gesture Recognition Results

Gesture recognition module was tested by a set of fixed predetermined gestures of the hands both in their fixed and moving form. In testing, the users made gestures in varying speeds, orientations and distance to the camera. The system was able to identify and identify the majority of gestures in real time. Gestures that were static like those involving counting fingers were highly recognized because they had definite shape-based characteristics. Swipes and directional movements were also found to be effective in dynamic gestures where patterns of motion were effectively defined.

B. Voice Control Results

To test the system, the system is subjected to varying lighting and noise conditions to test the accuracy and robustness of the system. Optimization measures like frame rate control and effective feature extraction are employed to make sure that it operates in real-time. The modular architecture enables one to test and enhance individual components.

C. Multimodal Interaction Results

Gesture Recognition and Voice Control System was also applied and tested to measure its performance, accuracy, reliability, and real time responsiveness. The findings indicate that the suggested system was successful in facilitating the natural and non-contact interaction between the user and the machine through hand gestures and voice commands. The system was also put to test in different conditions to determine its strength and its applicability in real life.

D. Real-Time Performance Results

Gesture recognition module was tested by a set of fixed predetermined gestures of the hands both in their fixed and moving form. In testing, the users made gestures in varying speeds, orientations and distance to the camera. The system was able to identify and identify the majority of gestures in real time. Gestures that were static like those involving counting fingers were highly recognized because they had definite shape-based characteristics. Swipes and directional movements were also found to be effective in dynamic gestures where patterns of motion were effectively defined.

E. Summary of Results

The outcomes of the Gesture Recognition and Voice Control System project show that the suggested system can fulfill the main purpose of providing the natural, non-contact, and intuitive human-computer interaction. It could recognize the hand gestures and voice commands in real time and translate them into suitable actions with great accuracy, which confirmed the possibility of multimodal interaction with the help of hardware that is readily available.

There were robust results with the gesture recognition module in the case of both static and dynamic gestures. Gestures in a static form depending on hand shape and the number of fingers got high recognition accuracy in ordinary lighting. Gestures of motion, which were dynamic that involved the movement of the hands were also well identified provided that the motion patterns were well defined. Though it is seen that there was a slight reduction in performance during low-light conditions or cluttered conditions, the preprocessing and hand-detection methods allowed ensuring acceptable accuracy.

The voice control unit was quite good in identifying predetermined voice commands especially in low-noise and moderately noisy settings. Sound control and system operations commands were perfectly performed with the least delay. The speech preprocessing method and noise reduction techniques enhanced the reliability, but the recognition accuracy reduced in noisy backgrounds.

The most important implication of the project is that gesture recognition and voice control are integrated effectively into a multimodal system. The joint utilization of the two input modalities was observed to increase system reliability and robustness. In cases where one of the input methods was influenced by environmental factors, the other was still operating well, so that there was no interruption of interaction.

This system was responsive in real-time with low latency and ensured good user experience as long as it was constantly operating. The users commented that the system was both user friendly and intuitive and they needed very little learning. On the whole, the findings indicate that the suggested system is feasible, effective, and applicable in a real-life context, and also provides the directions of future research, including sophisticated noise management and adaptive learning strategies.

F. Challenges of the Project

The creation of the Gesture Recognition and Voice Control System had a number of technical and practical difficulties, which impacted on the performance of the system and the complexity of its implementation. The gesture recognition in the diverse environmental conditions was one of the greatest challenges. Hand detection and segmentation accuracy were greatly affected by changes in lighting conditions, background clutter and camera angle. The complexity of the backgrounds and low-light situations also contributed to poor isolation of the hand region that decreased the recognition accuracy at times.

User variability was also another issue of concern. Gestures are varied among different users; their gesture speed, size of hands and style are different. It was challenging to design a system that would be consistent and not require much personalization and calibration to be used by all users. On the same note, the gestures can vary between cultures and thus it may not be easy to standardize.

Another significant problem was voice recognition when there was a lot of noise. There was background noise, overlapping speech and differences in accents and talking speed that influenced the accuracy of speech recognition. Noise reduction techniques were used but the performance was lower in high noise conditions.

Another problem was real-time processing constraints. The system was required to work with video and audio data at the same time and with small latency. In order to balance the accuracy and the computational efficiency of standard hardware, optimization was necessary.

There were also privacy and security issues because the constant use of cameras and microphones casts the issues of user trust. It was necessary to have minimal data storage and secure them.

On the whole, these issues made it inevitable to take into account the system design, the optimization of algorithms, and testing, and also outline the new areas of improvement.

VIII. Conclusion

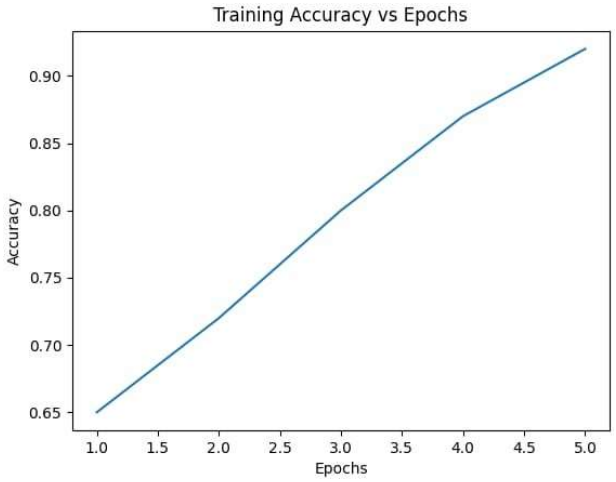

Fig. 4. This is my image caption

The Gesture Recognition and Voice Control System created in the given project is effective and intuitive to illustrate the natural human-computer interaction. The system eliminates the use of manual input devices, and allows hands-free operation by combining the hand gesture recognition system with voice-based control. The fundamental aims of the project are attained through the real-time performance, usability and affordability of the project in the form of readily available hardware like web camera and microphone. The system is able to identify both the fixed and live hand gestures accurately by the computer vision and voice commands by the speech

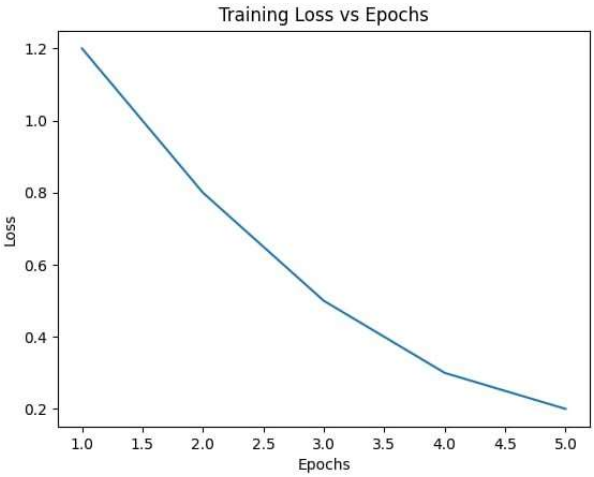

Fig. 5. This is my image caption

recognition procedure. The multimodality in inputting gestures and voice can also increase robustness, reliability, and make the system work despite the environmental conditions that may impair one of the input modalities. User testing attests the fact that the system is easy to use and needs minimum training which makes it applicable in practical application as smart homes, healthcare setups, assistive technologies, and interactive systems.

IX. Future Work

Even though the suggested system works efficiently, there are multiple improvements that can be considered in the further work to enhance performance and increase functionality. Among other things, there is the application of sophisticated methodologies of deep learning, including convolutional neural networks and recurrent neural networks, to enhance the efficiency of gesture recognition in complicated lighting and background environments. These models are able to acquire stronger features and deal with more variability in the user gestures. It is also possible to work on the improvement of voice recognition performance in very noisy settings in the future by including the use of complex techniques of noise cancellation and adaptive speech processing. The system should be made more accepting and accommodating to a global user base by adding support of multiple languages and accents to the system. Another direction of interest is personalization and adaptive learning. With time, the system can be improved to acquire the preferences of individual users, styles of gestures, and speech patterns, thus, increasing the level of accuracy and user satisfaction. Hand detection and tracking may be used further with integration with other sensors like depth cameras or wearable devices to make the process more accurate.

Acknowledgment

The authors would like to thank all contributors and supporters of this research for their invaluable assistance.

References

[1] N. K. Bhagat, Y. Vishnusai and G. N. Rathna, ”Indian Sign Language Gesture Recognition using Image Processing and Deep Learning.” 2019 Digital Image Computing: Techniques and Applications (DICTA), Perth, WA, Australia, 2019, pp. 1-8. doi: 10.1109/DICTA47822.2019.8945850.

[2] B. Natarajan et al., ”Development of an End-to-End Deep Learning Framework for Sign Language Recognition, Translation, and Video Generation,” in IEEE Access, vol. 10, pp. 104358-104374, 2022, doi: 10.1109/ACCESS.2022.3210543.

[3] T. M. Reddy, S. Abhishek, V. V. Kalyan, P. R. Varma and S. Sanapala, ”Sign Language Recognition Using OpenCV and Convolutional Neural Networks,” 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), Chennai, India, 2023, pp. 1-6, doi: 10.1109/RMKMATE59243.2023.10369046.

[4] Rasha Amer Kadhim, Muntadher Khamees A Real-Time American Sign Language Recognition System using Convolutional Neural Network for Real Datasets TEM Joumal. Volume 9, Issue 3, Pages 937-943, ISSN 2217-8309. DOI: 10.18421/TEM93-14, August 2020.

[5] G. A. Rao, K. Syamala. P. V. V. Kishore and A. S. C. S. Sastry. ”Deep convolutional neural networks for sign language recognition,” 2018 Conference Signal Processing and Communication Engineering Systems (SPACES), Vijayawada, India, 2018, pp. 194-197. doi: 10.1109/SPACES.2018.8316344.

[6] Thakur, Amrita, et al. ”Real time sign language recognition and speech generation.” Journal of Innovative Image Processing 2.2 (2020): 65-76.

[7] Dong, C., Leu, M. C., Yin, Z. (2015). American sign language alphabet recognition using Microsoft kinect. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 44-52).

[8] Singha, J., Das, K. (2013). Hand gesture recognition based on Karhunen-Loeve transform. arXiv preprint arXiv:1306.2599.

[9] Kang, B., Tripathi, S., Nguyen, T. Q. (2015, November). Real-time sign language fingerspelling recognition using convolutional neural networks from depth map. In 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR) (pp. 136-140). .

[10] C.K.M. Lee, Kam K.H. Ng, Chun-Hsien Chen, H.C.W. Lau, S.Y. Chung and Tiffany Tsoi, ”American sign language recognition and training method with recurrent neural network”, November 2020.

[11] Pei Xu, ”A real time hand gesture recognition and human-computer interaction system”, Proceeding of the Computer Vision and Pattern Recognition, 2017.

[12] V. Bhavana, G. M. Surya Mouli aras G. V. Lakshmi Lokesh, ”Hand Gesture Recognition Using Otsu’s Method”, 2017 IEEE International Conference Computational Intelligence and on Computing Research (ICCIC), pp. 1-4, 2017

[13] K. O. Rodriguez and G. C. Chavez, ”Finger Spelling Recognition from RGB-D Information Using Kernel Descriptor”, 2013 XXVI Conference on Graphics Patterns and Images, 2013

[14] Mukul Singh Kushwah. Manish Sharma. Kunal Jain and Anish Chopra, ”Sign language interpretation using pseudo glove”, Proceeding of International Conference on Intelligent Communication Control and Devices, pp. 9-18,2017.

[15] Yann LeCun, Yoshua Bengio and Geoffrey Hinton. Deep leaming. Nature, vol. 521, no. 7553, pp. 436-444, 2015.

[16] Alex Krizhevsky. Ilya Sutskever and E Hinton Geoffrey, ”Imagenet classification with deep convolutional neural networks”, Advances in neural information processing systems, pp. 1097-1105. 2012.

[17] Yann LeCun. Yoshua Bengio and Geoffrey Hinton, Deep leaming. Nature, vol. 521, no. 7553, pp. 436-444, 2015.

[18] Becky Sue Parton, ”Sign language recognition and translation: A multidisciplined approach from the field of artificial intelligence”, Journal of deaf studies and deaf education, vol. 11, no. 1. pp. 94-101. 2005

[19] Canzler, U., Dziurzyk, T.: Extraction of non-manual features for videobased sign language recognition. In: Proceedings of the IAPR Workshop on Machine Vision Applications, pp. 318-321 Nara, Japan (2002)

[20] S. Tamura and S. Kawasaki. Recognition of sign language motion images, In Pattern Recognition, volume 24, pages 343-353, 1988

[21] J. Ma, W. Gao, C. Wang, and J. Wu, A continuous Chinese Sign Language recognition system, International Conference on Automatic Face and Gesture Recognition, pages 428-433, 2000

[22] N. Tanibata, N. Shimada, and Y. Shirai, Extraction of hand features for recognition of sign language words, In Proc. Intl Conf. Vision Interface, pages 391-398, 2002.
[23] Yann LeCun, Yoshua Bengio and Geoffrey Hinton, Deep leaming. Nature, vol. 521, no. 7553, pp. 436-444, 2015.