International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 21 No. 4 (2025)

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 21, No. 4, 2025

Research Paper

SIGN LANGUAGE RECOGNITION USING CNN AND HAND GESTURES TRACKING

[1]K.MADHURIMA, [2]M.MANEESHA

[1]Assistant Professor, CSE, Tallapadmavathi College of Engineering, Somidi, Kazipet, Hanumakonda – 506003.Email-id: madhurimakotte@gmail.com.

[2]Research Scholar, H.no: 24UCID5807, CSE, Tallapadmavathi College of Engineering, Somidi, Kazipet, Hanumakonda – 506003,Email-id: maneeshamudigonda54@gmail.com.

ABSTRACT

Automatic hand gesture recognition from camera images is a compelling research area for developing intelligent vision systems. Sign language serves as the primary communication medium for individuals who are unable to speak or hear, enabling physically challenged people to express their thoughts and emotions effectively. In this work, we propose a novel scheme for sign language recognition aimed at accurately identifying hand gestures. Leveraging computer vision and neural networks, our system can detect gestures and convert them into corresponding text output. To tackle this problem, we introduce a 3D convolutional neural network (CNN) that automatically extracts discriminative spatio-temporal features from raw video streams without requiring handcrafted feature design. To enhance performance, multi-channel video streams—including color information, depth cues, and body joint positions—are used as input to the 3D CNN, allowing the integration of color, depth, and trajectory information for robust gesture recognition.

Index Terms: Sign Language Recognition, Hand Gesture Recognition, 3D Convolutional Neural Network (3D CNN), Spatio-Temporal Feature Extraction, Computer Vision, Depth Sensing, Human-Computer Interaction.

Received: 19-09-2025 Accepted: 22-10-2025 Published: 29-10-2025

1. INTRODUCTION

Sign language is one of the most widely used communication methods for hearing-impaired individuals, expressed through variations in hand shapes, body movements, and even facial expressions. Effectively capturing and interpreting these variations remains a challenging task due to the complexity of combining information from hand shapes and motion trajectories. This paper proposes an effective recognition model to translate sign language into text or speech, thereby helping hearing-impaired individuals communicate with others more seamlessly.

The primary technical challenge in sign language recognition (SLR) lies in developing robust descriptors to represent hand shapes and motion trajectories. Hand-shape representation involves tracking hand regions in video streams, segmenting hand-shape images from complex backgrounds in each frame, and recognizing gestures accurately [1, 2]. Motion trajectories require tracking key points over time and performing curve matching [3]. Despite extensive research in these areas, achieving satisfactory performance in SLR remains difficult due to variations and occlusions of hands and body joints [4]. Furthermore, integrating hand-shape and trajectory features effectively is nontrivial [5].

To address these challenges, we employ 3D convolutional neural networks (3D CNNs) to

naturally integrate hand shapes, action trajectories, and facial expressions [6]. Unlike traditional approaches that rely solely on color images as network inputs [1, 2], our model uses multiple visual streams—color images, depth images, and body skeleton images—simultaneously, all captured using Microsoft Kinect [7]. Kinect is a motion-sensing device that provides color and depth streams, and, with its public Windows SDK, allows real-time extraction of body joint locations (Fig. 1) [8]. Variations in color and depth at the pixel level provide crucial information for distinguishing sign actions, while temporal changes in body joints encode motion trajectories [9].

Using multiple visual modalities enables the CNN to focus on variations not only in color but also in depth and trajectory, effectively bypassing traditional challenges such as hand tracking, background segmentation, and manual feature design [10]. CNNs have the inherent capability to learn discriminative features automatically from raw data without prior knowledge [6]. Recently, 3D CNNs have shown promising results in video classification tasks [2, 4, 5]. Although training CNNs on large-scale video datasets can be time-consuming, real-time efficiency can still be achieved using parallel processing frameworks like CUDA [11].

In this work, we leverage 3D CNNs to extract spatio-temporal features from video streams for SLR. Unlike conventional methods that rely on handcrafted features to describe sign motion, our approach allows the network to learn motion and gesture patterns directly from raw video data, eliminating the need for manually designed descriptors [6, 12].

2. LITERATURE SURVEY

- Geoffrey E. Hinton, Ilya Sutskever, and Alex Krizhevsky [1] trained a large, deep convolutional neural network to classify 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into 1,000 different classes. On the test data, they achieved top-1 and top-5 error rates of 37.5% and 17.0%, which was considerably better than the previous state-of-the-art. The neural network, consisting of 60 million parameters and 650,000 neurons, includes five convolutional layers (some followed by max-pooling layers) and three fully connected layers with a final 1,000-way softmax output. To accelerate training, they used non-saturating neurons and an efficient GPU implementation of convolution. To reduce overfitting in fully connected layers, they employed the regularization method “dropout,” which proved highly effective. They further entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.
- Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei [2] demonstrated that convolutional neural networks (CNNs) are powerful models for image recognition and extended their application to large-scale video classification. Using a dataset of 1 million YouTube videos across 487 classes, they evaluated multiple approaches for extending CNN connectivity in the time domain to exploit spatio-temporal information. They proposed a multiresolution, foveated architecture to speed up training. Their best spatio-temporal networks showed significant performance improvements over strong feature-based baselines (55.3% to 63.9%), with modest gains over single-frame models (59.3% to 60.9%). Additionally, retraining the top layers on the UCF-101 Action Recognition dataset led to improved performance (63.3% up from 43.9%).
- Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner [3] discussed multilayer neural networks trained with back-propagation as a highly effective gradient-based learning technique. They showed that, given appropriate network architecture, gradient-

based learning algorithms can synthesize complex decision surfaces for classifying high-dimensional patterns, such as handwritten characters, with minimal preprocessing. Convolutional neural networks, specifically designed to handle 2D shape variability, were shown to outperform other techniques. They introduced Graph Transformer Networks (GTN) for global training of multimodule systems, demonstrating the advantage of global optimization in online handwriting recognition and document processing.

- H. Jhuang, T. Serre, L. Wolf, and T. Poggio [4] presented a biologically-motivated system for action recognition from video sequences. Their approach extends hierarchical feedforward architectures and neurobiological models of motion processing in the visual cortex. The system uses a hierarchy of spatio-temporal feature detectors: motion-direction sensitive units first analyze input sequences, producing position-invariant spatio-temporal features through successive processing stages. They found that sparse intermediate features outperform dense ones, and simple feature selection improves efficiency. Their approach achieved state-of-the-art results on multiple publicly available action datasets.

3.EXISTING SYSTEM

sign language into text or speech, facilitating communication between deaf-mute individuals and hearing people. This task has significant social impact but remains challenging due to the complexity and high variability of hand gestures and actions.

Existing methods for SLR primarily rely on hand-crafted features to describe hand motion and build classification models based on these features. However, designing reliable features that can generalize across diverse hand gestures is difficult, leading to limitations in accuracy and robustness.

DISADVANTAGES

- Heavy reliance on hand-crafted features, which may not capture all variations in hand gestures.
- Difficulty in tracking and segmenting hands accurately in real-time video streams.

4.PROPOSED SYSTEM

This paper aims to present a real time system for hand gesture recognition on the basis of detection of some meaningful shape based features like orientation, center of mass (centroid), status of fingers, thumb in terms of raised or folded fingers of hand and their respective location in image. The approach introduced in this paper is totally depending on the shape parameters of the hand gesture. It does not consider any other means of hand gesture recognition like skin color, texture because these image based features are extremely variant to different light conditions and other influences.

The system converts the Gestures video into simple words in English as well as make a sentence of that each word in English. The CNN process used in the video processing module gives the matched results. Based on the right match, the Sign Writing Image File is retrieved and stored in a folder. This folder served as the input to the Natural Language Generation Module.

ADVANTAGES

- The system processes gestures and provides immediate text output.
- CNN-based recognition ensures precise matching of hand gestures.

5.SYSTEM MODEL

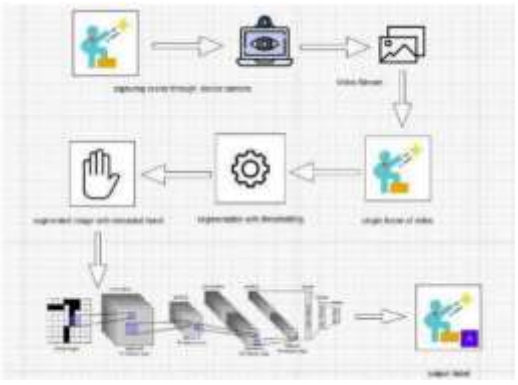


Fig.System Model

6.MODULES DESCRIPTION

1. Image Acquisition

Gestures are captured using a web camera. The OpenCV video stream records the entire signing duration, and individual frames are extracted from the stream. These frames are converted to grayscale images with a consistent dimension of 50 × 50 pixels, ensuring uniformity across the dataset for training and prediction.

2. Hand Region Segmentation & Hand Detection and Tracking

The captured frames are scanned to detect and segment hand regions. This preprocessing step enhances the visibility of hand gestures, making them more prominent for the recognition model. Effective segmentation and tracking significantly increase the accuracy of gesture prediction.

3. Hand Posture Recognition

The preprocessed hand images are fed into a Keras-based Convolutional Neural Network (CNN) model. The trained model generates predicted gesture labels with associated probabilities. The label with the highest probability is selected as the predicted gesture.

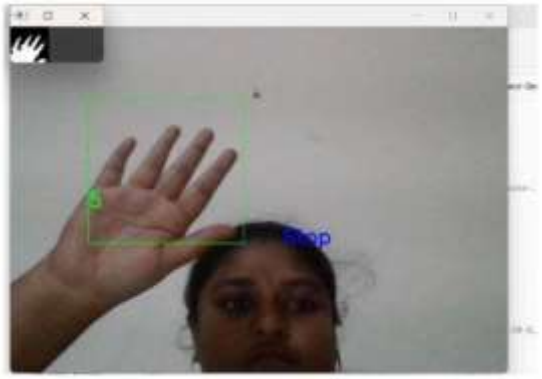
4. Display as Text & Speech

The recognized gestures are accumulated to form words.

RESULT

Run Application

Prediction



CONCLUSION

Nowadays, applications need several kinds of images as sources of information for elucidation and analysis. Several featuresare to be extracted so as to perform various applications. When an image is transformed from one form toanother suchas digitizing, scanning, and communicating, storing, etc. degradation occurs. Therefore , the output image has to undertake a process called image enhancement, which contains of a group of methods that seek to develop the visual presence of an image. Image enhancement is fundamentally enlightening the interpretability or awareness of information in images for human listeners and providingbetter input for other automatic image processing systems. Image then undergoes feature

extraction using various methods to make the image more readable by the computer. Sign language recognition system is a powerful tool to prepare an expert knowledge, edge detect and the combination of inaccurate information from different sources.

FUTURE ENHANCEMENT

Future Enhancement is being planned to further analyze and enhance the protocol usable for blind peoples because they need to communicate with normal persons

REFERENCES

1. L. Ku, W. Su, P. Yu, and S. Wei, "A real-time portable sign language translation system," 2015 IEEE 58th International Midwest Symposium on Circuits and Systems (MWSCAS), Fort Collins, CO, 2015, pp. 1-4, doi: 10.1109/MWSCAS.2015.7282137.
2. S. Shahriar et al., "Real-Time American Sign Language Recognition Using Skin Segmentation and Image Category Classification with Convolutional Neural Network and Deep Learning," TENCON 2018 – 2018 IEEE Region 10 Conference, Jeju, Korea (South), 2018, pp. 1168-1171, doi: 10.1109/TENCON.2018.8650524.
3. M. S. Nair, A. P. Nimitha, and S. M. Idicula, "Conversion of Malayalam text to Indian sign language using synthetic animation," 2016 International Conference on Next Generation Intelligent Systems (ICNGIS), Kottayam, 2016, pp. 1-4, doi: 10.1109/ICNGIS.2016.7854002.
4. M. Mahesh, A. Jayaprakash, and M. Geetha, "Sign language translator for mobile platforms," 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Udupi, 2017, pp. 1176-1181, doi: 10.1109/ICACCI.2017.8126001.
5. S. S. Kumar, T. Wangyal, V. Saboo, and R. Srinath, "Time Series Neural Networks for Real Time Sign Language Translation," 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, 2018, pp. 243-248, doi: 10.1109/ICMLA.2018.00043.
6. D. Kelly, J. Mc Donald, and C. Markham, "Weakly Supervised Training of a Sign Language Recognition System Using Multiple Instance Learning Density Matrices," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 41, no. 2, pp. 526-541, April 2011, doi: 10.1109/TSMCB.2010.2065802.
7. J. Jimenez, A. Martin, V. Uc, and A. Espinosa, "Mexican Sign Language Alphanumerical Gestures Recognition using 3D Haar-like Features," IEEE Latin America Transactions, vol. 15, no. 10, pp. 2000-2005, Oct. 2017, doi: 10.1109/TLA.2017.8071247.
8. M. Mohandes, M. Deriche, and J. Liu, "Image-Based and Sensor-Based Approaches to Arabic Sign Language Recognition," IEEE Transactions on Human-Machine Systems, vol. 44, no. 4, pp. 551-557, Aug. 2014, doi: 10.1109/THMS.2014.2318280.
9. D'Haro and F. Fernandez, "Speech into Sign Language Statistical Translation System for Deaf People," IEEE Latin America Transactions, vol. 7, no. 3, pp. 400-404, July 2009, doi: 10.1109/TLA.2009.5336641.
10. V. Lopez-Ludena, R. San-Segundo, R. Martin, D. Sanchez, and A. Garcia, "Evaluating a Speech Communication System for Deaf People," IEEE Latin America Transactions, vol. 9, no. 4, pp. 565-570, July 2011, doi: 10.1109/TLA.2011.5993744.