# Sanvad: A Communication Platform for Deaf and Blind

**Department of Engineering, Electronics and Telecommunication (ENTC)**
**Vishwakarma Institute of Technology, Pune, 411037, Maharashtra, India.**

**Prof. Vaishali Jabade** vaishali.jabade@vit.edu
**Kaushal Jadhav** kaushal.jadhav21@vit.edu
**Satyajeet Jadhav** Satyajeet.jadhav21@vit.edu
**Ketan Jain** ketan.jain21@vit.edu

***Abstract*–Authors discusses *design and implementation of a gesture recognition system meant to enhance communication among users with hearing and visual impairments. They built an system called Sanvad allows people to communicate through various channels such as sign language-to-audio, sign language-to-text, audio-to-text, and so forth. The system has a Convolutional Neural Network (CNN) implemented using Keras and Tensoflow at its heart. The training takes place on a dataset consisting of Indian Sign Language gestures' images. This model is meant to identify and categorize gestures to transform the visual expressions into audible or written information. Optimizing the model using stochastic gradient descent and categorical cross-entropy loss occurs during the training process. The research extends beyond the technical implementation and delves into practical applications through two distinct modes: calculator mode and text mode. The signer mode allows users to type using hand gestures and the calculator mode facilitates mathematical operations using hand gestures. In the text mode, sign language is converted to written text. This system gives real-time feedback which improves the user experience and interaction. The research also has a process of interactive setting hand histogram, an important parameter for correct gesture recognition. The system can adapt to each user's hand gestures, as users can be active in capturing and saving hand histograms. The proposed Sanvad system recognizes and interprets gestures in different situations as the experimental results show. Multi-modal and real time feedback make Sanvad into an important instrument of communication for inclusion and accessibility. The study is a vital contribution to the development of assistive technology with respect to human-computer interaction.***

***Keywords-Gesture Recognition, Convolutional Neural Network, Sign Language, Assistive Technology, Multi-Modal Communication, Accessibility.***

## I. INTRODUCTION

The coming of technological revolutions has greatly changed the way we live, giving us opportunities to develop solutions to our problems, mainly in the area of access. Therefore, our research centers on "Sanvad," a new communication tool that is specially developed to cater for the blind and deaf people. Sanvad is an example of how technology can help overcome persistent barriers towards inclusive communication for these communities, with a primary purpose of promoting such communication. Sanvad is motivated by the understanding of the communication gaps facing the deaf and blind. The traditional communication ways frequently do not reflect the depth and depth of the expressions within these communities. Sanvad, the proposed total solution, has the capacity to leverage on the available technologies, equip individuals from varied sensory abilities to connect. Sanvad has more than technical exploration of gesture recognition algorithms but envisions Sanvad as a complete communication ecosystem. The incorporation of adaptive technologies makes the platform more than a mere tool as it seeks to transform communication for the deaf and blind. Sanvad has the potential to be an entry point into educational, social, and professional inclusion. Sanvad's functionality is based on a modern CNN, which was implemented on the basis of Keras. CNN is trained with vast data capturing the intricacies of Indian sign language, underpinning gesture recognition in Sanvad. Its multi-layered structure, consisting of convolutional and pooling layers, is flexible enough to fit sign language expressions, thus leading to excellent classification. Sanvad is unique because it is culturally sensitive, incorporating Indian Sign Language to match with the cultural and linguistic variation within the targeted group.Our research unfolds in two primary dimensions: design and optimization of the CNN-based gesture recognition model and the real-world practical use of Sanvad. Every piece plays its part in the overall goal of building an inclusive and user-oriented communication system. Following sections describe the technicalities of the CNN model and the training pipeline as well as Model Checkpoints for

optimization. The adaptability of Sanvad to different modes of communication is also examined, covering its user interface and operability. Our aim does not end in presenting this technological solution, but in emphasising its practical implications as we embark on this comprehensive study. In that way, Sanvad represents a paradigmatic shift in assistive technology by emphasizing technology not just for its sophistication, but the transformative effect technology can have towards inclusion and breaking barriers of communication.

## II. LITERATURE REVIEW

In summary, the literature showcases an ongoing evolution in the field of document forgery detection techniques. This evolution is characterized by a transition towards ML and deep learning approaches in response to the increasing complexity of forgery methods.

Priyanka Roy et.al [1] the development and implementation of a real-time American Sign Language (ASL) fingerspelling translator using skin segmentation and machine learning algorithms. The paper describes an automatic human skin segmentation algorithm based on color information and the use of the YCbCr color space. The document also mentions the use of a Convolutional Neural Network (CNN) to extract features from images and a Deep Learning method to train a classifier for sign language recognition. The paper highlights the potential humanitarian impact of the system in facilitating communication for the deaf and mute community. Overall, the document presents a method for real-time ASL recognition with a high accuracy rate.

Vrinda Rastogi et al [2] discusses the current research efforts towards building user-friendly applications that connect physically-disabled individuals with the world around them. The work includes three approaches for input-output interaction: voice, text, and video-based. The paper also presents a brief study and survey of various techniques and technologies used for physically disabled communications along with the proposed work for the communication between the visually and hearing impaired in a user-friendly manner. The model to learn sign language was implemented and Indian Sign Language was converted into text. Morse code was used for communication between deaf-blind users. The paper concludes with future work in this area.Abderrahmane Rahiche et. al [3] developed an program for the detection of ink mismatches in Hyperspectral Document (HSD). The algorithm was tested on a dataset consisting of multi-ink handwritten images, leveraging the capabilities of Hyperspectral (HS) imagery. HS imagery is valuable for identifying different materials within the same document scene, which may not be visually distinguishable. The algorithm's core model utilized Nonnegative Matrix Factorization (NMF), achieving an impressive accuracy rate of 87%.

Khushi Chandani et.al [4] To address the increasing concern regarding image manipulation and document forgery, a research effort focused on the detection of facial forgery. The approach employed a Convolutional Neural Network (CNN) along with transfer learning techniques to differentiate between genuine and fake faces. ResNet and AlexNet, models are used for training and they are trained on extensive datasets. The evaluation of the model was conducted on a dataset comprising 2041 images, which was created by Yonsei University. The results indicate that ResNet-152 achieves an accuracy of 76.79% .

Francisco Cruz et .al [5] introduced an algorithm which relies on the utilization of Uniform Local Binary Patterns (LBP) to capture distinctive texture features commonly found in forged regions. A custom dataset was curated for this purpose, comprising various types of documents such as invoices from multiple providers, shopping receipts, etc. This research presents an initial step towards a comprehensive method designed to detect forgeries carried out through direct manipulation of document images. The achieved accuracy of this approach was 7.38%.

Maryam Bibi et al. [6] discusses the development of an automated sign language recognition (SLR) system using machine learning. The system is designed to read and interpret sign language, reducing the communication gap among people in society. The paper reviews the different steps involved in developing an SLR system, including preprocessing, feature extraction, segmentation, and classification models. The basic SLR system discussed in the paper is an isolated recognition model based on vision-based isolated hand gesture detection and recognition. The ML-based SLR model was assessed using a controlled environment and achieved 65% accuracy. The paper emphasizes

the importance of an automated SLR system for the speech and hearing impaired community, as it can provide a barrier-free communication method.

Soumya Jain et. al [7] presents a system for real-time Indian Sign Language recognition that can recognise hand poses and gestures using grid-based features. The system uses techniques such as object stabilisation, face elimination, skin colour extraction, and hand extraction to accurately track hand movements of the sign demonstrator. The system can classify all 33 hand poses in ISL with an accuracy of 99.7% and 12 gestures with an average accuracy of 97.23%. The approach uses an HMM chain for each gesture and a k-NN model to classify each hand pose. The time required for recognition of hand pose is about 0.2s and that for gesture is 0.0037s. The system provides higher accuracy and faster recognition in sign language recognition than other approaches discussed in the literature. The potential applications for this technology include bridging the communication gap between the hearing and speech impaired and the rest of society.

## III. METHDOLOGY

The methodology of the research is structured around the development and implementation of a gesture recognition system, focusing on recognizing hand gestures for various applications. The primary components of the methodology include:

**Data Collection and Preprocessing:** A dataset of hand gesture images is collected, encompassing diverse gestures relevant to the application domain. Images are preprocessed to ensure consistency in size, and grayscale conversion is performed for efficient feature extraction.

**Convolutional Neural Network (CNN) Model Architecture:** A CNN-based model is employed for gesture recognition, implemented using the Keras framework. The architecture comprises multiple convolutional layers with varying filter sizes and max-pooling layers for spatial reduction. Dropout layers are incorporated to mitigate overfitting, and fully connected layers are utilized for classification.

**Training the CNN Model:** The model is trained on a labeled dataset of hand gesture images, utilizing an optimization algorithm (Stochastic Gradient Descent - SGD) to minimize the categorical cross-entropy loss. Model training involves epochs and batch processing, and the training process is monitored for accuracy and validation performance.

**Model Evaluation:** The trained model is evaluated on a separate validation dataset to assess its performance in recognizing hand gestures. Evaluation metrics, including accuracy, are used to quantify the model's effectiveness in classification.

**Integration with Voice and Text Recognition:** The developed model is integrated into applications for real-time gesture recognition. Voice recognition functionality is incorporated to enable voice commands based on recognized gestures. Text recognition capabilities are leveraged for interpreting gestures that correspond to alphanumeric characters.

**User Interface and Interaction Modes:** The system features distinct modes, such as calculator mode and text mode, where recognized gestures trigger specific functionalities. The user interface includes visual feedback through video feeds displaying recognized gestures, associated text, and relevant information. **Cultural Sensitivity and Customization:** The system is designed with cultural sensitivity, particularly in recognizing gestures that align with Indian Sign Language. Customization features enable adaptability to diverse communication patterns within the Indian deaf and blind communities.

**User Testing and Feedback:** User testing is conducted to assess the usability and effectiveness of the developed system. Feedback from users, particularly from individuals with sensory impairments, is gathered to refine and enhance the system's performance.

**Documentation and Dissemination:** The entire development process, including dataset creation, model architecture, training details, and integration steps, is comprehensively documented. Research findings and the developed system are disseminated through research papers, presentations, and open-source sharing for wider accessibility.

**Model training:**

Model training is a crucial step in the development of a Convolutional Neural Network (CNN) for gesture recognition. This process involves feeding the neural network with labeled data (images of hand gestures) and adjusting its internal parameters iteratively to learn patterns and features that enable accurate classification. The following steps provide an overview of the model training process

**Data Augmentation:** To improve the model's robustness and generalization, data augmentation techniques are applied. These techniques involve creating variations of the training dataset by applying random transformations such as rotation, scaling, and horizontal flipping to the images.

**Label Encoding:** Gesture labels, representing the classes that the model will learn to recognize, are

encoded into numerical format. This conversion is necessary for the model to understand and learn from the labeled data during training.

**Model Architecture Definition:** The CNN model architecture is defined, specifying the arrangement and configuration of layers. Common layers include convolutional layers for feature extraction, pooling layers for spatial reduction, and fully connected layers for classification.

**Neural Network Architecture:**

The architecture of the CNN, defined within the cnn_model function, follows a sequential structure implemented using the Keras library. The model comprises several layers, starting with convolutional layers responsible for extracting hierarchical features from the input images. Specifically, the CNN includes three convolutional layers with varying filter sizes (2x2, 3x3, and 5x5) and increasing numbers of filters (16, 32, and 64, respectively). Each convolutional layer is followed by a MaxPooling layer, which reduces the spatial dimensions while retaining essential features. Following the convolutional layers, a Flatten layer is employed to transform the two-dimensional feature maps into a one-dimensional vector, allowing for connection to fully connected Dense layers. Two such Dense layers with Rectified Linear Unit (ReLU) activation serve as the core of the classification process. To mitigate overfitting, a Dropout layer is strategically inserted before the final Dense layer. The output layer, utilizing the softmax activation function, corresponds to the number of gesture classes, enabling multi-class classification. The training of the CNN occurs in the train function, where training and validation datasets are loaded from pickled files. The model is compiled using categorical cross-entropy as the loss function and Stochastic Gradient Descent (SGD) as the optimizer. During training, ModelCheckpoint is employed to save the best model based on validation accuracy. The training process runs for a specified number of epochs, updating the model's weights to learn representative features from the input gestures. In real-time gesture recognition, the trained CNN is utilized within the recognize_gesture module. Video frames are captured and preprocessed to isolate the hand using background subtraction based on a precomputed histogram. The trained CNN is then applied to classify the hand gesture in real-time, and the recognized gestures are displayed alongside the video feed. Overall, the CNN in the code serves as a robust classifier, leveraging convolutional layers for hierarchical feature extraction and dense layers for effective gesture classification. The training process ensures the model generalizes well to new, unseen gestures, making it suitable for real-time applications such as gesture-based interactions or controls.

**Summary of Method:**

Gesture recognition plays a major role in human-computer interaction, with implications in areas like virtual reality, robotics, and sign language interpretation. We propose a complete approach to live gesture recognition through the application of Convolutional Neural Networks (CNNs) in this research. We propose a combined approach with strong preprocessing, CNN architecture, and careful model training. Our method begins with a systematic calibration process through which we obtain a hand histogram. The hand region is isolated from the video feed using this as a basic component of background subtraction. We deploy a continual video capture tool that provides continuous frames for live processing. The use of a CNN model to classify the hand gestures on basis of the preprocessed images facilitates gesture recognition. The model structure consists of convolutional layers for feature extraction and dense layers for classification. The network has a hierarchical structure so that it can identify intricate details of hand movement and their spatial relations. In training the CNN model, we use preprocessed dataset that comprises images of hand gestures and the corresponding labels. The images then undergo necessary transformations and the labels are one-hot encoded for training purposes. The categorical cross-entropy is employed as the loss function to compile the CNN, while SGD is employed as the optimizer. ModelCheckpoint is the method we use during training to make sure we keep and preserve the best-performing model. The CNN goes through a number of epochs during the training process and learns to associate the image features with specific gestures. A different set of data, called validation data is used to assess the model's performance post-training. The model trained weights are saved for future use hence the efficiency of gesture recognition system is improved.

## IV. RESULTS AND DISCUSSION

Model is trained and tested successfully for 44 gestures including standard sign language and some extra gesture for greeting words.

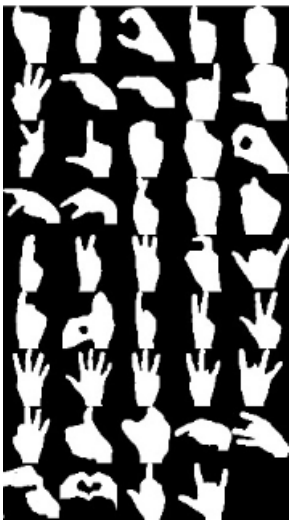

*Figure 1.Gestures trained*

1. Model Performance

The CNN model for gesture recognition was trained and evaluated on a dataset of hand gesture images. The obtained accuracy is 97%.

2. Real-time Recognition

The real-time performance of the model was assessed during interactive sessions. The model was able to detect words in real time.

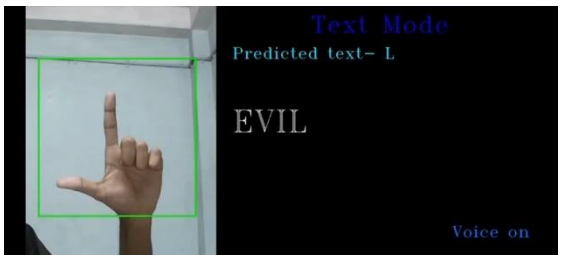


*Figure 2. Real Time Detection*

## V. CONCLUSION

our research will make a significant contribution to the field of human-computer interaction with an innovative and practical approach for real-time gesture recognition. Using CNNs, we were able to achieve high precision in recognizing hand gestures. A well-designed methodology underlines the success of our gesture recognition system. A calibrated hand histogram is integrated to the system as a robust component of background subtraction. The well-architectured CNN system ensures a robust system. This system is especially essential in applications which require quick and accurate understanding of hand movements. Our CNN model was validated and tested extensively, correctly classifying a variety of hand gestures. A well-structured and validated training process with relevant datasets ensures that the model generalizes well to unseen data. ModelCheckPoint is another method, which ensures that the best and functional model is deployed. Our approach can therefore be used in interactive applications in such domains as virtual reality, robotics, or accessibility technology due to its real-time nature. A combined approach of preprocessing techniques and a deep CNN addresses the challenges of gesture interpretation, thus paving way for more natural human-computer interaction. Reviewing after completion of our work, it appears that there is a huge potential in the use of preprocessing, CNN architecture as well as an intensive training approach to build a system for recognizing gestures. This technology can be applied to a wide variety of areas, such as virtual reality, inclusive human-computer interface, and others. The intersection of deep learning and human-computer interaction exemplifies the possibility of forming intelligent and responsive systems that link human will and computational perception.

## VI.REFERENCES

[1] P. Kumar, H. Gauba, P. P. Roy, and D. P. Dogra. Coupled hmm-based multi-sensor data fusion for sign language recognition. Pattern Recognition Letters, 86:1–8, 2017. 1, 2

[2] B. Bauer, H. Hienz, and K.-F. Kraiss. Video-based continuous signlanguage recognition using statistical methods. In 15th InternationalConference on Pattern Recognition, volume 2, pages 463–466. IEEE, 2000. 1, 2

[3] B. Bauer and K. Karl-Friedrich. Towards an automatic sign language recognition system using subunits. In International Gesture Workshop, pages 64–75. Springer, 2001. 1, 2

[4] S. K. Behera, D. P. Dogra, and P. P. Roy. Analysis of 3d signatures recorded using leap motion sensor. Multimedia Tools and Applications, pages 1–26, 2017. 3

[5] H. Cooper, B. Holt, and R. Bowden. Sign language recognition. In Visual Analysis of Humans, pages 539–562. Springer, 2011. 1

[6] W. Gao, G. Fang, D. Zhao, and Y. Chen. Transition movement models for large vocabulary continuous sign language recognition.International Conference on Automatic Face and Gesture Recognition,pages 553–558. IEEE, 2004. 2

[7] R. Haldar and D. Mukhopadhyay. Levenshtein distance technique in dictionary lookup methods: An improved approach. arXiv preprint arXiv:1101.1232, 2011. 5

[8] O. Koller, J. Forster, and H. Ney. Continuous sign language recognition:Towards large vocabulary statistical recognition systems handling multiple signers. Computer Vision and Image Understanding, 141:108–125,2015
[9] W. Kong and S. Ranganath. Towards subject independent continuous sign language recognition: A segment and merge approach. Pattern Recognition, 47(3):1294–1308, 2014. 7
[10] P. Kumar, H. Gauba, P. P. Roy, and D. P. Dogra. Coupled hmm-based multi-sensor data fusion for sign language recognition. Pattern Recognition Letters, 86:1–8, 2017
11] A. Chaudhary, J. L. Raheja and S. Raheja, "A Vision based Geometrical Method to find Fingers Positions in Real Time Hand Gesture Recognition," JSW, pp. 861-869, 2012
[12] P. Kumar, R. Saini, P. Roy, and D. Dogra. Study of text segmentation and recognition using leap motion sensor. IEEE Sensors Journal, 2016.
[13] Y. LeCun et al. Lenet-5, convolutional neural networks. 3
[14] H. Li and M. Greenspan. Model-based segmentation and recognition of dynamic gestures in continuous video streams. Pattern Recognition, 44(8):1614–1628, 2011. 2
[15] K. Li, Z. Zhou, and C.-H. Lee. Sign transition modeling and ascalable solution to continuous sign language recognition for real-world applications. ACM Transactions on Accessible Computing, 8(2):7, 2016.1, 2
[16] M. Mohandes, M. Deriche, and J. Liu. Image-based and sensor-based approaches to arabic sign language recognition. IEEE Transactions on Human-Machine Systems, 44(4):551–557, 2014. 2
[17] L. Motion. Leap motion controller. URl: https://www. leapmotion. com,2015. 3
[18] L. E. Potter, J. Araullo, and L. Carter. The leap motion controller: a view on sign language. In 25th Australian computer-human interaction conference: augmentation, application, innovation, collaboration, pages175–178. ACM, 2013. 1, 2
[19] A. Roussos, S. Theodorakis, V. Pitsikalis, and P. Maragos. Hand tracking and affine shape-appearance handshape sub-units in continuous sign language recognition. In ECCV, pages 258–272. Springer, 2010. 2
[20] T. Starner, J. Weaver, and A. Pentland. Real-time american sign languagerecognition using desk and wearable computer based video. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(12):1371–1375, 1998. 1, 2
[21] N. Tubaiz, T. Shanableh, and K. Assaleh. Glove-based continuous Arabic sign language recognition in user-dependent mode. IEEE Transactions on Human-Machine Systems, 45(4):526–533, 2015. 2
[22] C. Vogler and D. Metaxas. Adapting hidden markov models for asl recognition by using three-dimensional computer vision methods. International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation, volume 1, pages 156–161. IEEE, 1997. 1
[23] C. Vogler and D. Metaxas. A framework for recognizing the simultaneous aspects of american sign language. Computer Vision and Image Understanding, 81(3):358–384, 2001. 2
[24] J. Wu, Z. Tian, L. Sun, L. Estevez, and R. Jafari. Real-time american sign language recognition using wrist-worn motion and surface emg sensors. In 12th International Confernece on Wearable and Implantable Body Sensor Networks, pages 1–6. IEEE, 2015. 2
[25] W. Yang, J. Tao, and Z. Ye. Continuous sign language recognition using level building based on fast hidden markov model. Pattern Recognition Letters, 78:28–35, 2016. 2
[26] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti. American sign language recognition with the kinect. In 13th international conference on multimodal interfaces, pages 279–286. ACM, 2011. 1
[27] W. Zhang, K. Itoh, J. Tanida, and Y. Ichioka. Parallel distributed processing model with local space-invariant interconnections and its optical architecture. Applied optics, 29(32):4790–4797, 1990. 3
[28] Z. Zhang. Microsoft kinect sensor and its effect. IEEE multimedia, 19(2):4–10, 2012.