JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Communication System For Speech-Impaired People using Graph Neural Networks

Meena. T[1], Jaswanth Bojedla[2], Sri Tarak Ram Bandlamudi[3]
[1,2,3]Department of Computer Science and Engineering, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada, India.
Email: [1]meena@vrsiddhartha.ac.in, [2]jaswanthbojedla@gmail.com, [3] sritarakrambandlamudi@gmail.com

***Abstract*—Speech-impaired individuals are facing significant challenges in social interactions due to the lack of widespread knowledge of sign language among the general population. To address this issue, this research proposes a real-time communication system that leverages Graph Neural Networks (GNNs) to translate hand gestures into meaningful sentences, thereby bridging the communication gap between normal people and broader population effectively. The system captures live video of hand gestures using a webcam, pre-processes the images to enhance quality, and utilizes GNNs to classify graph-structured data, ensuring efficient and accurate gesture recognition. The recognized gestures are then converted into coherent sentences using Natural Language Generation (NLG), with the final out- put delivered as synthesized speech through a speaker. Unlike traditional hand gesture recognition systems that often translate gestures into isolated words, this system focuses on generating meaningful sentences, enabling natural and seamless communication. By integrating advanced techniques like GNNs for structured data processing and NLG for sentence formation, this proposed solution offers a significant improvement over existing methods, fostering inclusivity and enhancing the quality of life for speech-impaired individuals.**

***Index Terms*—Hand gesture recognition, graph Neural Networks, natural language generation, real-time communication, sign language translation, Indian Sign Language, assistive communication.**

I. INTRODUCTION

Communication is a fundamental human need, yet speech-impaired individuals often struggle to engage in meaningful interactions due to the widespread lack of sign language proficiency among the general population. This communication barrier significantly limits their ability to participate in social, educational, and professional settings, leading to feelings of isolation and exclusion. The challenge is further compounded by the reluctance of many to learn sign language, leaving speech-impaired individuals with few options to bridge this gap. According to a 2024 World Health Organization report, approximately 430 million people globally suffer from communication disorders, yet fewer than 10% have access to suitable assistive technologies. Recognizing the critical need for an inclusive and efficient solution, this research proposes a real-time communication system that leverages Graph Neural Networks (GNNs) to translate hand gestures into meaningful sentences, enabling more natural and accessible communi- cation. GNNs, known for their ability to process complex graph-structured data, ensure accurate recognition of intricate gestures, while a Natural Language Generation (NLG) module converts these recognized gestures into coherent sentences. The system employs a webcam to capture live hand gestures, preprocesses the data to enhance recognition accuracy, and delivers the output as synthesized speech. By integrating advanced machine learning and natural language processing methods, this work addresses the critical communication chal- lenges faced by speech-impaired individuals, fostering inclu- sivity and providing a practical, scalable solution to improve their quality of life.

II. RELATED WORKS

In [1],The authors implemented a system where high- resolution cameras were strategically placed near hive en- trances to capture real-time video footage of bees as they entered and exited the hive. The researchers were able to identify individual honey bees, marked and unmarked, through the use of image recognition and tracking algorithms. This allowed for the extraction of comprehensive behavioral data, such as feeding habits, hive congestion, and interactions be- tween bees. Convolutional neural networks (CNNs) were also used by the scientists for position estimation, which allowed for the accurate identification of bees bearing barcodes and the recognition of specific body sections of honey bees. The limitations of manual monitoring techniques were overcome by this automated monitoring system, which allowed for on- going, non-invasive observation and gave researchers a deeper understanding of bee behavior. Though this technology is excellent at tracking internal hive activities, it has a major drawback in that it isn't made to ward against outside dangers to the hive or shield the queen bee from predators. Hive security is not the major objective but behavioral monitoring is still the main goal.

In hand gesture recognition and voice conversion for deaf and dumb [2], the authors developed a hand gesture recogni- tion system aimed at facilitating real-time communication for individuals who are deaf or mute by translating hand gestures into text and audio. The approach integrated Google's Medi- aPipe framework with TensorFlow and OpenCV to detect hand

landmarks and classify gestures. The system consisted of three core components: capturing frames, detecting hand landmarks, and classifying gestures. For classification, the authors used a feed-forward neural network built with Keras, alongside an LSTM model to handle dynamic gestures and improve recognition accuracy. The system was trained to recognize ten gestures, achieving an accuracy rate of 95.7%. Pre-trained models were employed to extract features, which reduced the need for large-scale data collection. The results showed strong performance in varied lighting and backgrounds, but the confusion matrix indicated some mis-classification of similar gestures. Although the system performed well with static gestures, it faced challenges in handling dynamic sign languages. Despite this, the proposed system provided a practical solution for accessible communication by combining hand tracking with neural network-based gesture classification.

In a recent study Alphabet Recognition of Sign Language Using Machine Learning [3], the researchers set out to create an efficient system for recognizing sign language alphabets using advanced deep learning techniques. They utilized the Inception V3 Convolutional Neural Network (CNN) model and employed transfer learning on a comprehensive dataset of sign language gestures, with a focus on both Indian and Amer- ican Sign Language alphabets. This dataset, obtained from Kaggle, featured around 1200 images per letter, accounting for variations in hand positioning and orientation. The researchers applied several pre-processing steps, including segmentation and feature extraction, to enhance the image quality before feeding them into the model. As a result, the system achieved a high recognition accuracy of 98.99%, surpassing previous models, particularly in recognizing challenging letters such as ”C,” ”L,” and ”Y.” Additionally, the system was combined with a text-to-speech module, allowing for real-time conversion of recognized signs into spoken words. However, the system’s scope was limited to single-letter recognition, and future enhancements are suggested to expand its capabilities to complete words or full sentences. This work highlights the effectiveness of using the Inception V3 model and a robust dataset for developing sign language recognition sys- tems aimed at supporting individuals with hearing and speech impairments.

A network of interconnected sensors and devices within and around the hive is deployed as part of the technological implementation of the Internet of Things (IoT) concept for bee colony monitoring, as described by Anatolijs Zabasta [4]. These sensors continually gather information on the internal and exterior parameters of the hive. They are outfitted with features like temperature, humidity, sound, and weight tracking. The gathered information is subsequently sent in real- time using wireless communication protocols like Wi-Fi or LoRaWAN to a centralized cloud-based platform. Advanced analytics and machine learning algorithms are probably used on the cloud platform to evaluate the data and provide information on the behavior, health, and possible stresses of the hive. Actuators for automatic interventions, such as modifying environmental parameters or sending notifications to beekeepers in the case of irregularities, may also be incorporated into the Internet of Things system. This approach aims to improve beekeeping practices and contribute to the general health and sustainability of honey bee populations by developing and integrating hardware and software components into a strong and scalable Internet of Things infrastructure for colony monitoring. This method’s drawback is that, despite its extensive monitoring capabilities, it might not be able to handle any threats from predators or the outside environment that could jeopardize the colony’s health and safety.

In order to collect and capture the acoustic signals generated by the colony, S. Ferrari investigated the placement of audio sensors strategically inside the hive. He specifically concen- trated on noises that are typical of pre-swarming activity [5]. After that, the gathered audio data is sent to a centralized processing unit, where it may be combined with machine learning algorithms intended to discern between swarming noises and other hive sounds. To improve accuracy, a dataset of recognized swarming occurrences may be used to train the algorithm. By continuously analyzing the acoustic data, patterns that point to an approaching swarm may be identified early and beekeepers can receive timely notifications. By enabling proactive management methods during the crucial swarming phase, this non-invasive and automated strategy en- hances colony retention and overall beekeeping efficiency. The swarming monitoring system makes use of cloud computing, wireless connectivity, data storage, alarm systems, and aural sensors. This method’s small drawback is that it could miss certain important hive activities or environmental elements that could alter swarming behavior, which could have an impact on the monitoring system’s overall efficacy.

III. METHODOLOGY

The methodology begins by converting hand gesture images into structured graph representations using OpenPose, which detects key hand joints such as fingertips, knuckles, and the wrist. These key points are then used to construct a skeletal graph, which is processed by a Graph Neural Network to classify gestures into corresponding words. The recognized words are refined into grammatically correct and coherent sentences using TextBlob, ensuring linguistic accuracy. The generated text is then converted into speech using text-to- speech technology, enabling verbal communication. This struc- tured process effectively bridges hand gestures and spoken language, facilitating seamless interaction. The methodology consists of 4 major steps.They are :

A. Graph Structure Generation

This step focuses on converting hand gesture images into graph-based representations to allow the GNN (Graph Neural Network) to interpret the data effectively. The purpose is to simplify the data by highlighting key structural features like joint positions, which helps in reducing complexity and enhancing model performance. OpenPose, a pose estimation

method, is used to extract key points such as finger tips, knuckles, and the wrist. These points serve as nodes, while the links between them form edges, creating a skeletal representation of the hand. The process takes a preprocessed hand gesture image as input and produces a graph structure as output, which will be fed into the GNN in the subsequent phase of the hand gesture-to-speech project.

The skeletal structure of hand gestures was obtained using OpenPose, which identifies key points such as the wrist, knuckles, and finger tips from input images. These key points were connected according to the anatomical layout of the hand, forming a graph representation. The method involved determining the (x, y) coordinates of each joint and connecting them with edges to outline the skeletal framework. OpenPose employs part affinity fields (PAFs) to accurately predict both the positions of key points and the connections between them, enabling the generation of precise graph structures for further analysis with GNNs.

Part Affinity Fields (PAFs):

$$f(x) = \sum_{i=1}^{N} \delta(x - x_i) \cdot \mathbf{v}_i \quad (1)$$

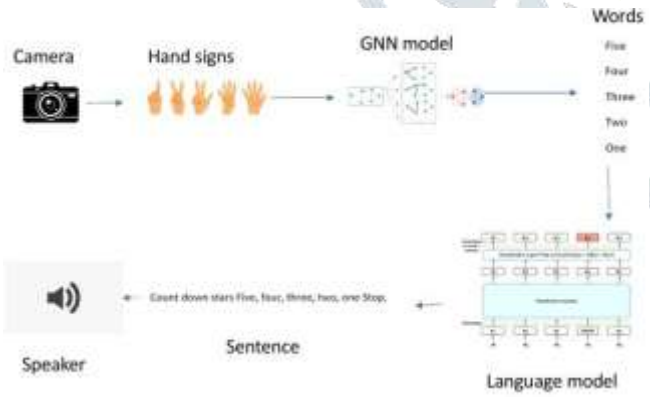

Fig. 1. Proposed System Workflow

B. Graph-Based Hand Gesture Classification

In this phase, a Graph Neural Network (GNN) was trained to classify hand gestures by processing the graph representations created earlier. The GNN takes in a graph where nodes represent the key points of the hand, and edges capture the relationships between these points. Graph convolutional layers were used to update node features by aggregating information from neighboring nodes, allowing the model to capture the spatial structure of the hand gesture. At the output layer, a softmax activation function was employed to predict the class of the gesture. During training, labeled graph data was used, with each gesture corresponding to a specific label. The model's objective was to minimize the cross-entropy loss, which was optimized using backpropagation. This process allowed the GNN to learn the underlying patterns in the gesture graphs, enabling it to classify new, unseen gestures accurately.

Graph Convolution Update:

$$h_i^{(k+1)} = \sigma\left(W^{(k)} \cdot \sum_{j \in N(i)} c_{ij} h_j^{(k)} + b^{(k)}\right) \quad (2)$$

where:

- $h_i^{(k+1)}$ is the updated feature of node i at the $(k+1)$-th layer,
- $N(i)$ represents the set of neighboring nodes of node i,
- $W^{(k)}$ is the learnable weight matrix at layer k,
- c_{ij} is a normalization factor for edge ij,
- $b^{(k)}$ is the bias term at layer k,
- σ is an activation function.

Softmax Function:

$$p(y = k \mid x) = \frac{\exp(z_k)}{\sum_{j=1}^{K} \exp(z_j)} \quad (3)$$

where:

- $p(y = k \mid x)$ is the probability of input x belonging to class k,
- z_k is the score (logits) for class k,
- K is the total number of possible gesture classes.

Cross-Entropy Loss:

$$L = -\sum_{k=1}^{K} y_k \log(p(y = k \mid x)) \quad (4)$$

where:

- y_k is the true label for class k (1 if the class is correct, 0 otherwise),
- $p(y = k \mid x)$ is the predicted probability for class k.

C. Gesture-to-Sentence Conversion

In this phase, the individual words derived from the Graph Neural Network (GNN) are transformed into complete sentences using TextBlob, a Python-based library for natural language processing. Once the GNN classifies each gesture into a corresponding word, these words are passed through TextBlob to form coherent sentences. TextBlob provides several features such as part-of-speech tagging, word inflection, and noun phrase extraction, which help in generating grammatically correct and contextually appropriate sentences. The words are processed in sequence, allowing TextBlob to apply linguis- tic rules to structure them into fluent sentences. This step effectively converts isolated gesture words into well-formed sentences, making it possible to output them as either text or speech for seamless communication. TextBlob handles tasks like correcting the tense of verbs, adjusting plural forms, and ensuring the overall structure aligns with standard grammar rules. The model also accounts for syntactic consistency, ensuring that subject-verb agreement is maintained throughout the sentence. After generating a grammatically sound sentence, the output is then ready for further processing, such

as converting it into speech or displaying it as text. This entire process enables a seamless transition from isolated gesture words to meaningful sentences that can be used for communication, providing an essential step toward making gesture-based interaction more intuitive and natural.

D. Speech Generation

In this final step, the sentences generated from the gesture-to-sentence conversion process are transformed into speech or audio output using Text-to-Speech (TTS) technology. This technology takes the generated text and converts it into spoken language. The process involves feeding the text into a TTS system, which synthesizes the corresponding phonetic and prosodic features, such as intonation, rhythm, and pitch, to produce a natural-sounding audio output. Popular TTS system, Google Text-to-Speech, employ advanced deep learning techniques to create clear and expressive speech that mimics human vocal patterns. By converting the generated sentences into speech, this step allows the system to provide a voice-based interface for users, making it easier for individuals to interact with the system using spoken language. This conversion is particularly useful in making sign language communication more accessible to a wider audience.

IV. ALGORITHM

The presented methodology is a comprehensive system aimed at enabling speech-impaired individuals to communicate effectively by translating hand gestures into spoken language. It includes several stages, starting with the extraction of skeletal key points from hand gestures and their conversion into graph-based representations. These graphs are then pro- cessed using a GNN model to classify the gestures, followed by sentence formation through natural language generation (NLG). Finally, a text-to-speech (TTS) module converts the generated sentences into audible speech. This approach en- hances accessibility and provides a practical solution for real- time communication in various everyday situations.

Gesture Recognition and Speech Generation System -1 0 textbf

INPUT: Hand gesture images captured from a camera or dataset

OUTPUT: Generated speech or sentence

Step 1: Set parameters for gesture classification model (e.g., GNN) and speech generation

Step 2: Data Pre-processing

Step 2.1: Apply OpenPose for skeleton detection and extraction of key points (e.g., wrist, knuckles, fingertips)

Step 2.2: Convert detected key points into graph structures, representing joints as nodes and connections as edges

Step 3: Model Training for Gesture Recognition

Step 3.1: Train the GNN model on the gesture dataset to classify hand gestures

Step 3.2: ReLU activation function: $f(x) = \max(0, x)$

Step 3.3: Apply graph convolutional layers to learn gesture-specific features

Step 3.4: Loss function for classification: $L = -\sum y \cdot \log(\hat{y})$ (cross-entropy loss)

Step 4: Sentence Generation

Step 4.1: Input isolated words (e.g., from GNN output) to the NLG system

Step 4.2: Use TextBlob or alternative method for sentence construction

Step 5: Speech Generation

Step 5.1: Use Text-to-Speech (TTS) system to convert generated sentence to speech

Step 5.2: Output speech via speaker or audio device

V. RESULTS AND DISCUSSION

The model was evaluated using a variety of hand gesture inputs, with the results demonstrating robust performance in both gesture classification and sentence formation. The Graph Neural Network (GNN) model achieved a classification accuracy of 90%, showcasing its capability to accurately interpret the hand gestures. Once classified, the words were processed by TextBlob, which successfully generated sentences in most cases. However, some minor errors were observed in the sentences due to the inherent limitations of TextBlob in handling complex sentence structures and context. Despite these small inaccuracies, the system effectively translated gesture-based words into coherent sentences.

These results indicate that the combination of the GNN for gesture classification and TextBlob for sentence generation provides an efficient and reliable solution for gesture-to-text conversion. The model's high accuracy and the overall quality of the generated sentences affirm the feasibility of this approach for real-time applications, enabling effective communication through gesture-based inputs.

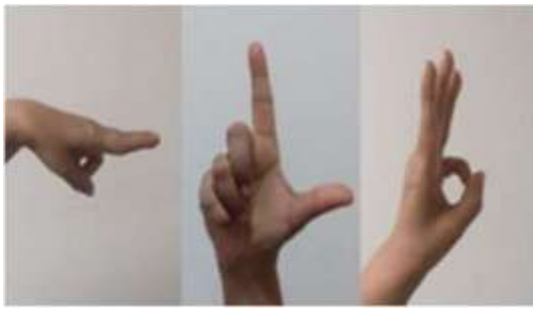

Fig. 2. Hand Gestures Input

This system provides a valuable solution for individuals with speech impairments by converting sign language gestures

Fig. 3. Sentence Output

Fig. 4. Overall Conversion Mechanism

into sentences and generating speech. In social settings, it enables non-verbal individuals to communicate directly with others without the need for an interpreter. In professional contexts, the system allows speech-impaired people to participate in conversations and meetings, supporting their inclusion. Additionally, in public environments, it enhances their ability to interact with services, increasing independence. By offering real-time gesture-to-speech translation, the system helps elim- inate communication barriers, making interactions smoother. In essence, it promotes a more inclusive environment for individuals with speech impairments.

VI. CONCLUSION AND FUTURE WORK

The project, ”Communication System for Speech-Impaired People Using Graph Neural Networks,” successfully demonstrated its capability to classify hand gestures and convert them into coherent sentences with 90% accuracy. By employ- ing Graph Neural Networks (GNN) for gesture recognition and TextBlob for sentence formation, the system effectively bridges the gap between sign language gestures and speech, enabling speech-impaired individuals to communicate. This solution enhances accessibility in both social and professional environments, fostering greater inclusivity for individuals with speech impairments.

Looking ahead, future developments will focus on creating a more advanced language model to generate sentences from isolated words produced by the GNN. This model will aim to provide better compatibility and accuracy compared to TextBlob. Additionally, the project will work on improving the Text-to-Speech (TTS) system to produce more natural- sounding speech. To further enhance user experience, a mobile application will be developed, facilitating real-time gesture- to-speech interaction, thereby supporting speech-impaired in- dividuals in a wide range of contexts and contributing to a more inclusive society.

VII. REFERENCES

[1] Navali, Samarth Kolachalam, Jyothirmayi Vala, Vanraj. (2021). Sentence Generation Using Selective Text Prediction. *Computacio´n y Sistemas*, 23. 10.13053/cys-23-3-3252.

[2] Palivela, Hemant. (2022). Optimization of paraphrase generation and identification using language models in natural language processing. *International Journal of Information Management Data Insights*, 1, 100025. 10.1016/j.jjimei.2021.100025.

[3] Zhao, H.; Chen, H.; Ruggles, T.A.; Feng, Y.; Singh, D.; Yoon, H.-J. (2024). Improving Text Classification with Large Lan- guage Model-Based Data Augmentation. *Electronics*, 13, 2535. https://doi.org/10.3390/electronics13132535.

[4] Akshaya, R., Sindhu, Daniel. (2023). Hand Gesture to Speech Trans- lation using Deep Learning. *International Journal of Advanced Re- search in Science, Communication and Technology (IJARSCT)*, 3(13). DOI:10.48175/568.

[5] Rianti, Afika Widodo, Suprih Ayuningtyas, Atikah Hermawan, Bima. (2022). Next Word Prediction Using LSTM. *Journal of Information Technology and Its Utilization*, 5. 10.56873/jitu.5.1.4748.

[6] Rahim, M. A., Shin, J., Islam, M. R. (2021). Dynamic Hand Gesture Based Sign Word Recognition Using Convolutional Neural Network with Feature Fusion. In *2021 IEEE 2nd International Conference on Knowledge Innovation and Invention (ICKII)*, Seoul, Korea (South), pp. 221-224. doi:10.1109/ICKII46306.2019.9042600.

[7] Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Mian, A. (2023). A comprehensive overview of large language models. *arXiv preprint* arXiv:2307.06435.

[8] Miah, Abu Saleh Musa Hasan, Md. Al Shin, Jungpil. (2023). Dynamic Hand Gesture Recognition Using Multi-Branch Attention Based Graph and General Deep Learning Model. *IEEE Access*, PP, 1-1. 10.1109/AC- CESS.2023.3235368.

[9] Vaniukov, S., Vaniukov, S. (n.d.). How to Build a Large Language Model: Step-by-Step Guide. *Softermii*. Retrieved from https://www.softermii.com/blog/how-to-build-a-large-language-model- step-by-step-guide.

[10] The Hindu. (n.d.). India has five million peo- ple with communication disabilities. Retrieved from https://www.thehindu.com/news/national/karnataka/india-has-five- million-people-with-communication-disabilities/article65065563.

[11] Khattak, Faiza Jeblee, Serena Pou-Prom, Chloe Abdalla, Mohamed Meaney, Christopher Rudzicz, Frank. (2019). A survey of word em- beddings for clinical text.

[12] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Yu, P. S. (2020). A Comprehensive Survey on Graph Neural Networks. *IEEE Trans- actions on Neural Networks and Learning Systems*, 32(1), 4-24. doi:10.1109/TNNLS.2020.2978386.

[13] Kipf, T. N., Welling, M. (2017). Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[14] Yan, S., Xiong, Y., Lin, D. (2018). Spatial Temporal Graph Convolu- tional Networks for Skeleton-Based Action Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

[15] Yuan, H., Yu, H., Wang, J., Li, K., Zhang, S., Yu, Y., Lin, Y. (2022). Graph Neural Networks for NLP: A Survey. *Computational Linguistics*, 48(1), 1–41. doi:10.1162/coli a 00436.

[16] Wang, X., Girshick, R., Gupta, A., He, K. (2020). Non-local Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelli- gence*, 42(4), 1037-1051. doi:10.1109/TPAMI.2019.2958754.

[17] Zhou, J., Cui, G., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M. (2020). Graph Neural Networks: A Review of Methods and Applications. *AI Open*, 1, 57-81. doi:10.1016/j.aiopen.2021.01.001.

[18] Chung, J., Gulcehre, C., Cho, K., Bengio, Y. (2014). Empirical Evalua- tion of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv preprint* arXiv:1412.3555.

[19] Dogan, Y., Yildiz, A., Yildirim, H. H. (2023). Real-Time Dynamic Hand Gesture Recognition using CNN and LSTM Fusion for Human- Computer Interaction. *Journal of Ambient Intelligence and Humanized Computing*. doi:10.1007/s12652-023-04129-x.

[20] Wang, Y., Sun, H., Li, W., Li, T., Li, Y. (2021). Diverse Beam Search for Natural Language Generation with Sentence-Level Diversity. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pp. 1050–1061.