International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI:10.55041/ISJEM05745
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Spiking Neural Network Model for Hand Gesture Interpretation and Text Output

1.**Merugu Manish Kumar, 2.Mogili Ravindar, 3.Jagannatham Krishnahitha, 4.Bollaram Shiva, 5.Anasuri Sriya**
Department of CSE (Artificial Intelligence & Machine Learning)
Jyothishmathi Institute of Technology and Science, Karimnagar, Telangana, India manishmerugu13@gmail.com,
jagannathamkrishnahitha@gmail.com, shivabollram@gmail.com, anasurisriya@gmail.com

***Abstract*—**This project presents a Spiking Neural Network (SNN)–based system for real-time hand gesture interpretation and text generation. The system uses MediaPipe to extract 21- point hand landmarks from live video input. These features are encoded into spike trains using Time-To-First-Spike encoding and processed by an SNN model to recognize both static and dynamic hand gestures. Recognized gestures are converted into meaningful text outputs to support natural human–computer in- teraction and assistive communication. Compared to conventional deep learning approaches, the proposed system is lightweight, energy-efficient, and suitable for real-time execution on resource- constrained devices. The results demonstrate the effectiveness of SNNs for gesture-based communication applications.

***Index Terms*—**Spiking Neural Networks, Hand Gesture Recog- nition, Neuromorphic Computing, MediaPipe, Deep Learning, Text Output

1.INTRODUCTION

Hand gesture recognition has become an important area of research due to its applications in human–computer in- teraction, assistive communication, and touch-free systems. Gestures provide a natural and intuitive way for humans to convey information without relying on speech or physical contact. This is especially useful for individuals with hearing or speech impairments, as well as for modern interactive systems where traditional input devices are not practical.

Most existing gesture recognition systems are built using deep learning models such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). While these models achieve high accuracy, they often require large com- putational resources, high power consumption, and longer processing times, which limits their use in real-time and edge- based applications. Additionally, many systems focus mainly on static gestures and struggle to efficiently capture temporal gesture dynamics.

Spiking Neural Networks (SNNs) offer a biologically in- spired alternative that processes information in the form of discrete spikes over time. This makes SNNs well suited for modeling temporal patterns while maintaining low energy consumption. In this project, MediaPipe is used to extract precise hand landmarks from live video input, which are then encoded into spike trains using Time-To-First-Spike encoding. These spike-based representations are processed by an SNN to recognize hand gestures efficiently. The proposed system not only identifies hand gestures but also converts them into meaningful text outputs, enabling effective gesture-to-text communication. By combining real-time hand tracking with spiking neural processing, the system aims to provide a lightweight, responsive, and practical solution for gesture-based interaction and assistive communication.

2.BODY OF THE PAPER

2.1 RELATED WORK

Hand gesture recognition has been extensively explored in the fields of computer vision and human–computer interaction. Early research primarily relied on handcrafted features combined with classical machine learning techniques such as Support Vector Machines (SVMs), k-Nearest Neighbors (k-NN), and Hidden Markov Models (HMMs). While these methods performed adequately in constrained environments, their robustness was limited when applied to real-world scenarios involving variations in lighting, background, and hand orientation.

With the rise of deep learning, Convolutional Neural Networks (CNNs) became the dominant approach for gesture recognition tasks. CNN-based models demonstrated strong performance in recognizing static hand gestures from image data. To address temporal information in dynamic gestures, researchers later integrated Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks with CNN architectures. Although these hybrid models improved recognition accuracy, they introduced significant computa- tional overhead and higher power consumption, restricting their suitability for real-time and edge-based systems.

Recent studies have shifted towards landmark-based and skeleton-based representations to reduce dependency on raw image input. Hand landmark extraction frameworks such as MediaPipe enabled reliable detection of key hand joints, allowing gesture recognition systems to focus on geometric and motion features. This approach improved robustness to environmental changes while lowering computational complexity.Spiking Neural Networks (SNNs) have emerged as a bi- ologically inspired alternative to traditional neural networks, particularly for temporal data processing. SNNs operate using discrete spikes, enabling efficient modeling of time-dependent patterns with reduced energy consumption. Encoding tech- niques such as Time-To-First-Spike and rate coding have been proposed to adapt conventional gesture data for spiking architectures. However, existing work on SNNs has largely focused on simple classification tasks, with limited exploration of real-time gesture-to-text communication systems.

Based on the reviewed literature, there is a clear need for a lightweight and real-time gesture recognition frame- work that combines landmark-based input with spiking neural processing. The proposed work builds upon these studies by integrating MediaPipe-based hand tracking with an SNN model to achieve efficient hand gesture interpretation and meaningful text generation.

2.2 SYSTEM DESIGN

The proposed system is designed to recognize hand gestures from real-time video input and convert the recognized gesture into corresponding text output. The overall system pipeline is divided into four major stages: video acquisition, landmark extraction, spike encoding with SNN-based classification, and text generation.

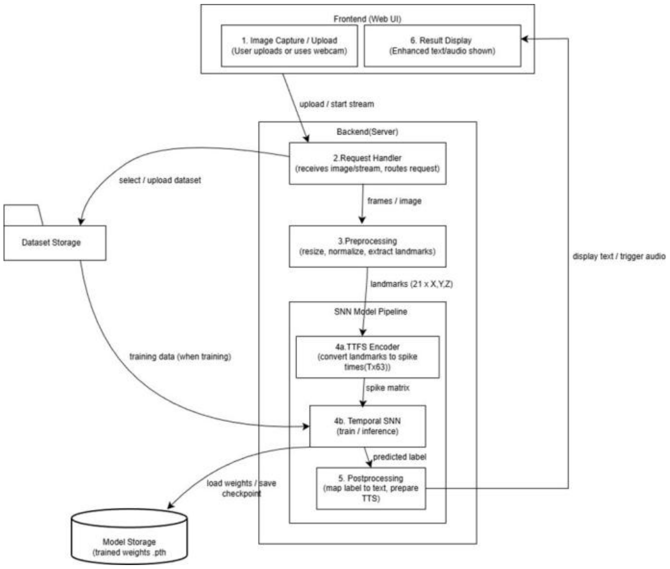


Fig. 1. System Architecture

A. *Proposed System Architecture*

Input Acquisition: The system captures live video frames using a webcam or built-in camera module. Each frame is processed sequentially to ensure real-time performance. The captured frames are resized and normalized to maintain stable processing speed and uniform input conditions.

Hand Landmark Detection: For accurate gesture interpretation, the system uses MediaPipe Hand Tracking to detect the hand region and extract 21 key landmark points. These landmarks represent the joints and finger positions of the hand. Landmark-based representation reduces dependency on background conditions and improves robustness under varying illumination and cluttered environments.

Spike Encoding: Since Spiking Neural Networks process information as spike events, the extracted landmark features are converted into spike trains using Time-To-First-Spike (TTFS) encoding. In this method, stronger input activations generate spikes earlier, enabling efficient temporal coding.

This encoding converts continuous landmark coordinates into time-based spike patterns, making the data suitable for SNN processing and reducing computational overhead com- pared to conventional deep learning methods.

Spiking Neural Network (SNN) Classifier: The spike-encoded input is fed into the Spiking Neural Network, which consists of:

- Input layer (spike trains from landmarks)
- Hidden spiking layers (feature learning)
- Output layer (gesture class prediction)

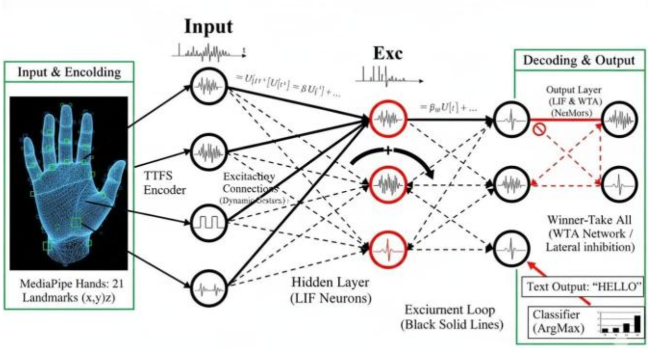


Fig. 2. Internal architecture of the Spiking Neural Network (SNN) classifier used for hand gesture recognition and text output generation.

The SNN learns temporal patterns of both static and dynamic gestures. Due to event-driven processing, the SNN provides low latency and reduced power consumption, making it suitable for real-time gesture recognition applications.

Gesture to Text Generation: After classification, the predicted gesture label is mapped to its corresponding text output. The system generates meaningful words or sentences (instead of only raw gesture class names). This improves usability, especially for assistive communication applications such as sign language interpretation.

Output Interface: The final output text is displayed on the user interface in real time. Optionally, the system can also provide speech output using text-to-speech modules for enhanced interaction.

B. *Modules*

- Gesture Input Acquisition
- Hand Landmark Detection (MediaPipe)
- Spike Encoding
- Spiking Neural Network Classification
- Text Generation / Output Display

2.4 IMPLEMENTATION

This project was implemented as a complete real-time hand gesture recognition system using a Spiking Neural Network (SNN). The main goal of the implementation is to capture live hand gestures through a camera, detect hand landmarks, convert the landmark data into spike-based input, and finally classify the gesture and generate the corresponding text output. The entire implementation is divided into multiple modules so that each stage works clearly and efficiently.

A. Tools and Libraries Used

The project was developed using Python along with the following libraries:

- **Python**: Used as the main programming language for building the entire workflow.
- **MediaPipe Hands**: Used for detecting the hand and extracting 21 landmark points from each frame.
- **OpenCV (cv2)**: Used to access the webcam, capture real- time video frames, and display output on screen.
- **NumPy**: Used for handling landmark arrays, preprocess- ing operations, and saving processed data.
- **PyTorch**: Used for implementing and training the spiking neural network model.
- **Scikit-learn**: Used for dataset splitting and validation support during training.

B. Data Collection and Input Handling

The first step in implementation is capturing gesture input. For this, the webcam is accessed using OpenCV and frames are collected continuously. The system processes the frames one by one to maintain real-time response. While collecting data, multiple samples of each gesture were recorded so that the model can learn gesture variations like different hand positions, speed changes, and small movement differences.

C. Hand Landmark Extraction using MediaPipe

After capturing the frames, MediaPipe Hand Tracking is used to detect the hand in each frame and extract 21 landmark points. These landmarks represent important hand joints and finger positions. Each landmark provides (x, y, z) coordinate values. This landmark-based approach is very useful because it reduces dependency on background and lighting, and focuses only on the hand structure.

For every frame, these landmark points are converted into a feature vector which becomes the input representation for further processing.

D. Spike Encoding using TTFS

Since SNN models do not work like normal neural net- works, the extracted landmark values cannot be given directly. Therefore, spike encoding is performed. In this project, **Time- To-First-Spike (TTFS)** encoding is used. In TTFS encoding, stronger input values produce spikes earlier, which makes the data suitable for temporal learning in SNN.

This encoding step converts continuous landmark coordinate values into spike train patterns, which helps the model learn gesture dynamics efficiently with lower computational cost.

E. SNN Model Implementation

The spike-encoded input is then passed to the Spiking Neural Network classifier implemented in PyTorch. The model contains:

- An **input layer** that receives the spike trains
- One or more **hidden spiking layers** for feature learning
- An **output layer** that predicts the gesture class

The main advantage of using SNN here is that it can learn temporal information from gestures while providing quick response and efficient computation. This makes it suitable for real-time dynamic gesture recognition.

F. Training Process

For training the model, the recorded gesture samples are preprocessed and converted into spike inputs. A dataset loader is used to load training samples in batches. The model is trained using the Adam optimizer with multiple epochs. During training, validation is also performed to check whether the model is learning correctly and to reduce overfitting.

After training, the best performing model weights are saved and later used for real-time gesture recognition.

G. Real-Time Gesture Recognition and Text Output

In real-time testing mode, webcam frames are taken continuously and a buffer of frames is maintained for dynamic gesture recognition. Each frame is processed through the same pipeline: landmark extraction, spike encoding, and SNN classification. The predicted gesture label is mapped into the corresponding text output.

Finally, the generated text output is displayed on the inter- face in real time. This helps the system act as an assistive communication tool, where gesture-based inputs can be converted into meaningful text.

2.5 RESULTS AND DISCUSSIONS

This section presents the results obtained from the implemented Spiking Neural Network based hand gesture recognition system and discusses the overall performance of the model in real-time conditions. The system was tested using multiple gesture samples collected through webcam input under different lighting conditions and hand orientations.

A. Dataset and Testing Setup

To evaluate the system, gesture samples were collected as dynamic sequences. Each gesture sample consists of multiple frames, and for every frame, 21 hand landmark points were extracted using MediaPipe. These landmarks were encoded using TTFS spike encoding and then passed to the trained SNN model.

The testing was performed on a standard system with webcam input. The output was displayed in real time to verify whether the predicted gesture matches the performed gesture.

B. Model Prediction Results

The trained Spiking Neural Network was able to classify the hand gestures and generate the corresponding text output successfully. During testing, the model produced correct predictions for most gestures when the hand was clearly visible to the camera.

The following observations were made:

- The system performed well for gestures with clear finger separation and stable motion.
- Gesture recognition was accurate when the input gesture sequence duration was consistent.
- Real-time output generation was smooth and responsive due to event-driven spike processing.

C. Discussion

From the obtained results, it is observed that landmark- based input representation plays a major role in improving recognition performance, because it reduces the effect of background clutter and illumination changes. Also, TTFS encoding helped in representing landmark values as temporal spike patterns, making the data suitable for SNN processing.

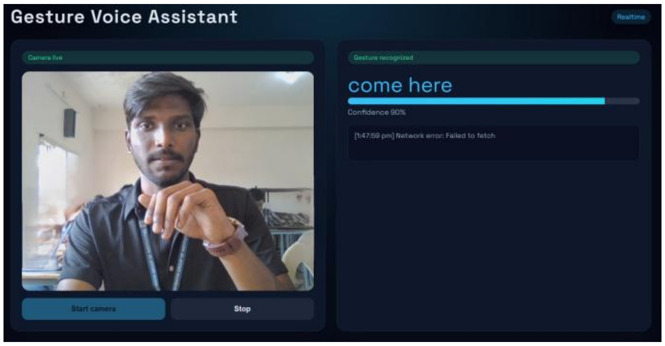
However, a few limitations were also observed:

- If the hand moves very fast or goes out of the camera frame, landmark extraction becomes inaccurate and prediction may fail.
- Some gestures with similar hand shapes may be misclas- sified if the temporal motion pattern is not distinct.
- Performance may reduce in low lighting conditions where hand detection becomes unstable.

Overall, the system demonstrates that combining MediaPipe hand tracking with TTFS-based spike encoding and SNN classification is an efficient approach for dynamic hand gesture interpretation. The results confirm that the proposed model can be used for real-time gesture-to-text applications with low latency and computational efficiency.

TABLE I
PERFORMANCE COMPARISON OF CNN AND SNN-BASED GESTURE RECOGNITION MODELS

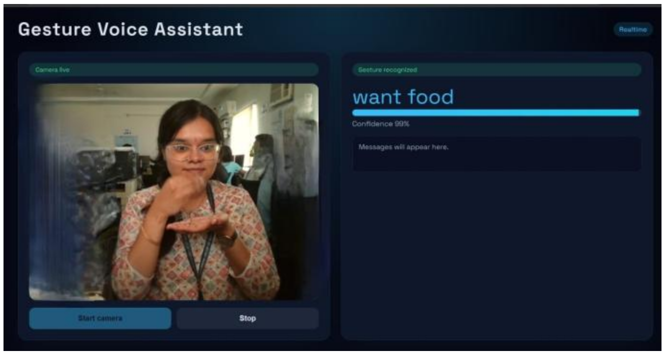
Model	Dataset	Gesture Type	Acc. (%)	Remarks
CNN-based HGR	Static RGB gesture dataset	Static	97.8	High accuracy CNN; computationally heavy, limited temporal modeling
Event-Based SNN	DVS-Gesture / DVS-GC	Dynamic	96.5	Energy-efficient SNN; depends on event-camera hardware
Radar-based SNN	Radar gesture dataset	Dynamic	98.0	Privacy-safe and accurate; requires costly radar hardware
Proposed SNN (This Work)	ASL + custom dynamic dataset (MediaPipe)	Dynamic	98.0	TTFS-based SNN; real-time input with text output



come here

Confidence 90%

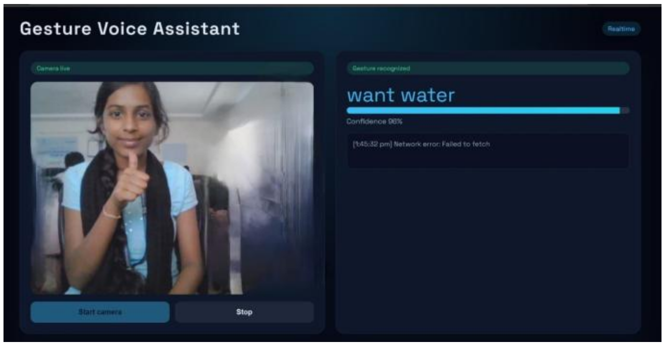
[1:47:59 pm] Network error: Failed to fetch



want food

Confidence 99%

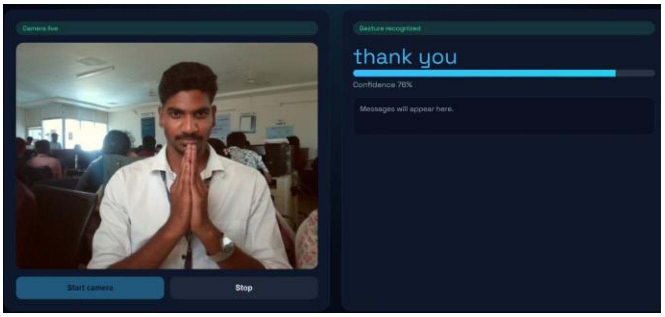
Messages will appear here.



want water

Confidence 96%

[1:45:32 pm] Network error: Failed to fetch



thank you

Confidence 76%

Messages will appear here.

3. CONCLUSION AND FUTURE WORK

This paper presented a Spiking Neural Network (SNN) based approach for real-time hand gesture interpretation and text output generation. The proposed system captures live video input, extracts robust hand landmark features using MediaPipe, converts these features into temporal spike patterns using Time-To-First-Spike (TTFS) encoding, and performs gesture classification using an SNN classifier. The recognized gesture is then mapped to meaningful text output, enabling an effective gesture-to-text communication workflow.

From the implementation and experimental observations, it is evident that landmark-based representation improves stability by reducing the influence of background variations, while spike-driven processing helps in achieving low-latency inference suitable for real-time applications. The developed system successfully demonstrates the feasibility of combining neuromorphic computing principles with computer vision to support gesture-driven interaction and assistive communication.

In future, this work can be extended by increasing the ges- ture vocabulary and training on larger datasets collected from multiple users to improve generalization across different hand shapes, lighting conditions, and motion variations. Further improvements can also be achieved by supporting continuous gesture sequences for sentence-level interpretation, enhancing temporal learning using hybrid deep learning models, and integrating optimized deployment as a web or mobile application with real-time speech output for better accessibility.

REFERENCES

[1] Vicente-Sola, Alex *et al.*, ”Spiking Neural Networks for event-based action recognition: A new task to understand their advantage.” Neuro- computing 611 (2025): 128657.

[2] Tsang, Ing Jyh *et al.*, ”Radar-based hand gesture recognition using spiking neural networks.” Electronics 10.12 (2021): 1405.

[3] Chen, Xuena *et al.*, ”Sign language gesture recognition and classification based on event camera with spiking neural networks.” Electronics 12.4 (2023): 786.

[4] Kirkland, Paul *et al.*, ”Unsupervised spiking instance segmentation on event data using STDP features.” IEEE Transactions on Computers 71.11 (2022): 2728-2739.

[5] Davies, Mike *et al.*, ”Advancing neuromorphic computing with loihi: A survey of results and outlook.” Proceedings of the IEEE 109.5 (2021): 911-934.

[6] W. Maass, Networks of spiking neurons: the third generation of neural network models, Neural Netw. 10 (9) (1997) 1659–1671.

[7] M. Davies, A. Wild, G. Orchard, Y. Sandamirskaya, G.A.F. Guerra, P. Joshi, P. Plank, S.R. Risbud, Advancing neuromorphic computing with loihi: A survey of results and outlook, Proc. IEEE 109 (5) (2021) 911–934.

[8] A. Amir, B. Taba, D.J. Berg, T. Melano, J.L. McKinstry, C. di Nolfo, T.K. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza, J.A. Kusnitz, M.V. DeBole, S.K. Esser, T. Delbru¨ck, M. Flickner, D.S. Modha, A low power, fully event-based gesture recognition system, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 7388–7397.

[9] A. Kugele, T. Pfeil, M. Pfeiffer, E. Chicca, Efficient processing of spatio-temporal data streams with spiking neural networks, Front. Neurosci. 14 (2020) 439.

[10] P. Kirkland, D. Manna, A. Vicente-Sola, G. Di Caterina, Unsupervised spiking instance segmentation on event data using STDP features, IEEE Trans. Comput. (2022).

[11] P. Lichtsteiner, C. Posch, T. Delbruck, A 128× 128 120 dB 15 s latency asynchronous temporal contrast vision sensor, IEEE J. Solid- State Circuits 43 (2) (2008) 566–576,

[12] T. Delbru¨ck, B. Linares-Barranco, E. Culurciello, C. Posch, Activity-driven, event based vision sensors, in: Proceedings of 2010 IEEE International Symposium on Circuits and Systems, 2010, pp. 2426–2429.

[13] J. Lee, T. Delbru¨ck, M. Pfeiffer, P.K.J. Park, C.-W. Shin, H. Ryu, B.-C. Kang, Real-time gesture interface based on event-driven processing from stereo silicon retinas, IEEE Trans. Neural Netw. Learn. Syst. 25 (2014) 2250–2263.

[14] Y. Bi, Y. Andreopoulos, PIX2NVS: Parameterized conversion of pixel-domain video frames to neuromorphic vision streams, in: 2017 IEEE International Conference on Image Processing (ICIP), 2017, pp. 1990–1994.

[15] D. Gehrig, M. Gehrig, J. Hidalgo-Carri’o, D. Scaramuzza, Video to events: Recycling video datasets for event cameras, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3583–3592.