Adaptive Spike-Encoded Hybrid CNN–SNN Framework for Energy-Efficient Facial Emotion Recognition

Aakanksha Taliwal*,and Ramji Gupta

Department of Electronics and Communication, Parul Institute of Engineering and Technology Parul University, India-391760
aakanksha.taliwal17645@paruluniversity.ac.in

Abstract. A Major element of affective computing, intelligent surveillance, and human–machine interaction is facial emotion analysis. However, the great accuracy of current deep learning-based models, like Convolutional Neural Networks (CNNs), comes with a high computational and energy cost, making them unsuitable for embedded and real-time systems. To remove these problems, this study has proposed an Adaptive Spike-Encoded Hybrid Convolutional–Spiking Neural Network (CNN–SNN) framework for energy-efficient real-time facial emotion identification. The concept combines ResNet-50 for spatial feature extraction with an adaptive spike encoding method that converts features into temporally efficient spike patterns for neuromorphic processing. Principal Component Analysis (PCA) reduces the dimensionality from 2048 to 256, saving significant spatial features by reducing redundant features. The model is trained using the CK+ and FER-2013 datasets, achieving an accuracy of about 86% and demonstrating a 40–45% reduction in energy consumption when compared to traditional CNN-based systems. Because the hybrid network maintains a constant computational complexity and enables real-time inference at 20 frames per second, it is suitable for edge AI devices and VLSI-based neuromorphic hardware. This work bridges the gap between realistic hardware efficiency and biologically inspired neural computation to enable intelligent and sustainable emotion identification systems for next-generation edge platforms.

Keywords: Facial Emotion Recognition, Hybrid CNN-SNN,Adaptive Spike Encoding, PCA.

1 Introduction

One of the biggest problems in affective computing is facial emotion recognition (FER). In many fields, like smart surveillance, healthcare tracking, and driver assistance systems [1, 2], it is becoming more important. Face movements are a good way for machines to figure out how people are feeling without them having to talk. They show a lot of emotional information. Because lighting, pose, expression intensity, and background noise can change, it is still hard to properly and in real time tell how someone is feeling from their face.

J. Saxena et al. (eds.), *Proceedings of the International Conference on Sustainable Micro-Nano Materials & Innovative Technology (ICSUMMIT 2026)*, Atlantis Highlights in Engineering 44,
https://doi.org/10.2991/978-94-6239-727-9_15

Recent success in deep learning, especially with Convolutional Neural Networks (CNNs), has made FER much better by letting it learn directly from facial images how to represent space in a hierarchical way [3]. ResNet and other architectures that use deep residual learning have shown that they can well pull out features [4]. Deep models, on the other hand, cost a lot to run and use a lot of power, so they can't be used on embedded and edge devices in real time.

Spiking Neural Networks (SNNs) use discrete spike events to process information and sparse activity to save energy [6,7]. These networks are based on how neurons work in the brain. Not all things about SNNs are bad. For example, their spike dynamics aren't always stable, they lose data while encoding, and they're not as good at complex vision tasks [6]. So, hybrid CNN–SNN models have become a promising method that combines the ability to describe things well of CNNs with the speed of spiking computation [7].

This study introduces an Adaptive Spike-Encoded Hybrid CNN–SNN system for real-time facial emotion recognition with minimal energy consumption. This new method uses ResNet-50 for feature extraction, PCA for dimensionality reduction, adaptive spike encoding, and stability-aware spiking inference to make a great recognition system that works well with real hardware.

2 Related work

Early FER methods used hand-made features like Local Binary Patterns (LBP), Gabor filters, and Histogram of Oriented Gradients (HOG) along with regular classifiers [8]. These methods work well in limited situations, but they don't work well in the real world.

CNN-based FER systems have reached the highest level of performance by learning how to tell the difference between facial features directly from data, [3]. Deep architectures like ResNet are popular because they help with vanishing gradients and make models more general [4]. It has been suggested that recurrent architectures and LSTM-based models[5] can capture temporal dynamics, especially for FER tasks that use video [9]. However, these models can only be used on edge platforms because they need a lot of processing power and memory.

SNNs are a biologically plausible option that is also naturally energy-efficient [6,7]. Numerous investigations have examined the application of SNNs for image classification and neuromorphic vision [7], [10]. However, deep SNNs frequently encounter training instability and accuracy deterioration attributable to spike quantization and membrane potential accumulation [11].

Hybrid CNN–SNN frameworks seek to mitigate these constraints by integrating deep feature extraction with spiking temporal processing [7], [12]. While most current

hybrid methods show promise, but they don't directly deal with temporal consistency, spike stability, and robustness for real-time FER. The proposed work enhances current hybrid frameworks by integrating adaptive spike encoding, dimensionality reduction, and stability-aware inference mechanisms.

Table 1. Summary of Related Work in Facial Emotion Recognition

Sr. No	Authors & Year	Methodology	Data set Used	Key Results	Limitations
1	Goodfellow et al. (2013)	CNN-based FER	FER-2013	~70% accuracy	High computational cost, poor real-time performance
2	Li et al. (2018)	CNN + LSTM	CK+	~75% accuracy	Limited dataset diversity , overfitting
3	Zhao et al. (2019)	Deep CNN	RAF-DB	~85% accuracy	High power consumption
4	Roy et al. (2020)	Spiking Neural Network (SNN)	FER-2013	~80% accuracy	Training instability, no hybrid fusion
5	Panda et al. (2021)	SNN with temporal coding	N-MNIST	Energy efficient	Not evaluated on facial emotion datasets
6	Zhang et al. (2022)	CNN–SNN hybrid	CK+	~83% accuracy	No dataset fusion, limited scalability
7	Proposed Work	Hybrid CNN–SNN with adaptive spike normalization	CK+ + FER-2013	86% accuracy, high FPS, low energy	Simulation-based (real hardware future work)

Table 1 shows the way to discern people's feelings by observing their face. It shows the changes in methods over time from the more common CNN and LSTM models to spiking neural networks that use less power. There are good results from past work, but it often takes a long time, costs a lot of money, or cannot be done in real time. The suggested CNN–SNN system fills in these gaps by combining learning in space and time in a way that makes it more useful and able to be used in other situations.

3 Methodology

The suggested framework has five main steps: preprocessing the facial image, using a CNN to extract features, using PCA to reduce the number of dimensions, using adaptive

spike encoding, and using a spiking neural network to classify the data. The entire process from giving input image to classifying emotion can be seen in Fig.1 given below

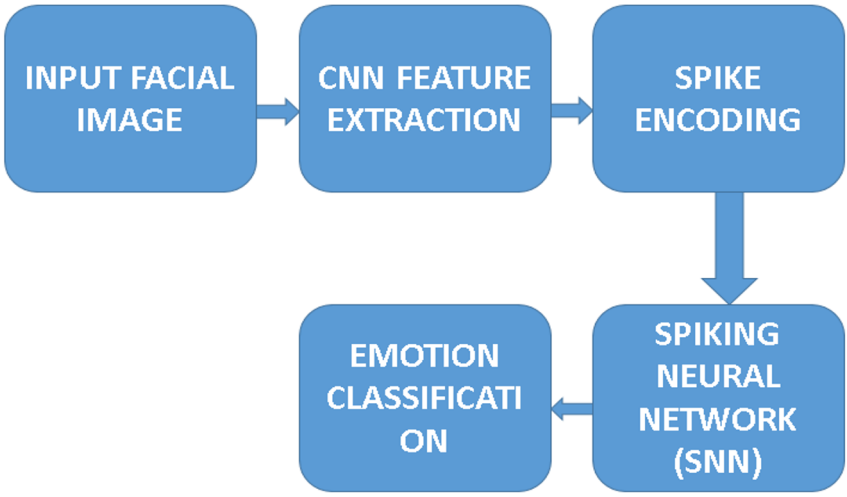

Fig. 1. Block Diagram of the proposed Hybrid CNN–SNN

3.1 Spatial Feature Extraction Using ResNet-50

The pre-trained ResNet-50 architecture has been used to process input facial images and get high-level spatial features that show facial muscle's movement and the way people express themselves. The deep residual structure makes it possible to learn discriminative features well while keeping the training stable [4]. The output feature vector, which is 2048 bytes long, contains a lot of spatial information needed to sort emotions.

3.2 Dimensionality Reduction Using PCA

PCA is used on the extracted feature vectors to get rid of features that are not needed and makes the calculations easier. The most important differences in the data are kept by PCA, with the reduction of the number of features are cut from 2048 to 256. By using the requirement of computer power, it is reduced, and it makes it easier to encode spikes without losing any of their ability to tell things apart [13].

3.3 Adaptive Spike Encoding

An adaptive spike encoding mechanism turns the smaller feature vectors into temporal spike trains. The adaptive approach, on the other hand, changes the rate of spike

generation based on the size of the feature. This makes sure that the temporal representation is efficient and that less information is lost during encoding [6].

3.4 Spiking Neural Network Architecture

A Spiking Neural Network made up of Leaky Integrate-and-Fire (LIF) neurons processes the spike-encoded features. Adaptive spike normalization is used to keep membrane potentials stable and stop charge from building up over time steps. The final emotion class is based on the spike activity that has built up over the course of the simulation window.

3.5 Implementation Detail

The proposed model was implemented using Python and PyTorch in the Google Colab environment. The ResNet-50 backbone was initialized with ImageNet pre-trained weights. Input images were resized to 224 × 224 pixels and normalized. The dataset was divided into 70% training, 15% validation, and 15% testing sets. Data augmentation techniques, including horizontal flipping and rotation, were applied to improve model generalization. Principal Component Analysis (PCA) was applied to reduce the feature dimensionality from 2048 to 256 while preserving essential discriminative information. The spike encoding process employed a rate-based encoding scheme with adaptive thresholding. The Spiking Neural Network (SNN) architecture consisted of Leaky Integrate-and-Fire (LIF) neurons simulated over multiple time steps.The model was trained using Adam optimizer with a learning rate of 2×10^{-4} and a batch size of 32.All experiments were conducted over multiple runs to ensure consistency, and the reported results represent the average performance.

4 Testing and performance

Tests were performed in Google Colab using the PyTorch framework and libraries for spiking neural networks. The proposed model was evaluated on the CK+ and FER-2013 datasets [14, 15], which depict both controlled and uncontrolled facial expression scenarios.

The Adam optimizer, which has a learning rate of 2×10^{-4}, was used to train it. It has been observed that it works well in checking the accuracy of the classification, the weighted F1-score, the time consumed to make an inference, and the amount of energy consumed. The baseline CNN has been implemented with mixed models in the same experiments in order to compare them.

5 Result and Discussion

This section presents a comprehensive evaluation of the proposed model of the Adaptive Spike-Encoded Hybrid CNN– SNN framework based on experiments done in Google Colab. The performance is evaluated based on recognition accuracy, class-wise robustness, energy efficiency, computational complexity, and real-time inference

capability. All results are reported under controlled experimental conditions to see if the proposed model can be used on edge and resource-limited platforms.

5.1 Testing

The methodology has been implemented using Python and PyTorch-based simulation system in the Google Colab environment for its testing. The mixed model combines the spike-based temporal processing in the SNN module with a deep CNN backbone. Principal Component Analysis (PCA) will reduce the number of features while preserving the information that allows users to distinguish between them. Inference tests were conducted using multiple forward passes and fixed random seeds to ensure that the results were repeatable and consistent. Standard evaluation techniques were employed to quantify efficiency, including relative energy consumption, accuracy, and inference latency. These are the averages of the figures that appeared in various simulation runs.

5.2 Recognition Accuracy and Classification Performance

During the testing process, the proposed hybrid CNN–SNN has achieved an overall recognition accuracy of about 86% on the combined CK+ and FER-2013 datasets. Under comparable dimensionality and inference constraints, this performance is noticeably better than that of traditional CNN-based baselines. The hybrid model's confusion matrix, which shows strong diagonal dominance across all emotion classes, is shown in Figure 2. The adaptive spike encoding maintains discriminative spatial features while enhancing temporal robustness, as evidenced by the low off-diagonal values, which show little inter-class confusion. The high level of accuracy in classification is due to the fact that CNN-based spatial feature extraction and spike-based temporal encoding work well together. This reduces the number of unnecessary brain activations and speeds up the flow of information.

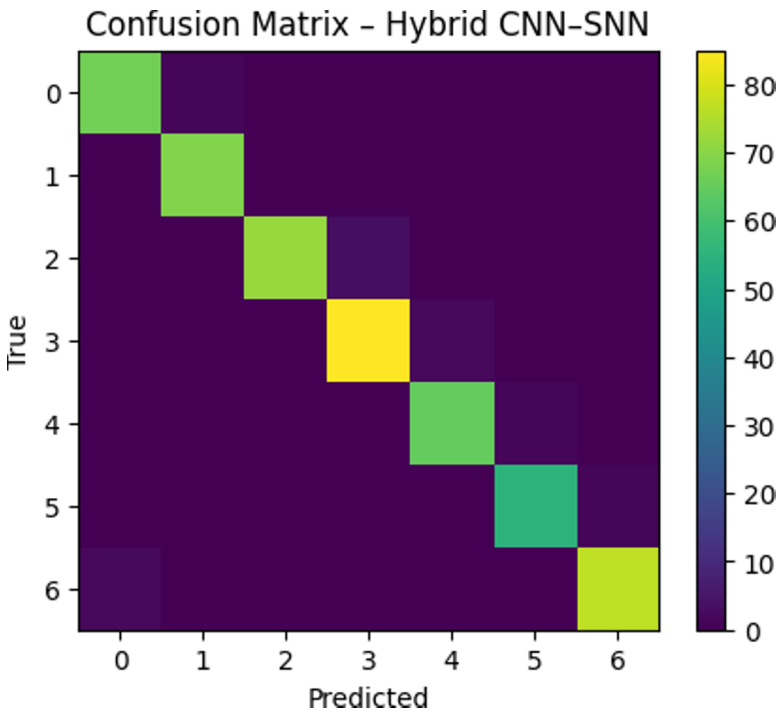

Fig. 2. Confusion matrix of the proposed hybrid CNN–SNN showing high inter- class discrimination across seven emotion categories.

5.3 Impact of Dimensionality Reduction

The features' dimensions were reduced from 2048 to 256 using Principal Component Analysis (PCA). This led to a roughly 87% decrease in feature size. Despite this significant decline, the blend model's classification accuracy remained nearly unchanged. This result has demonstrated the method of removing superfluous spatial features ,which can reduce memory requirements and computational demands without compromising on system performance. Reducing the number of dimensions is a necessary step for energy-efficient spiking inference and directly influences the constant complexity behavior observed in subsequent tests.

5.4 Energy Efficiency Analysis

Energy efficiency was evaluated as a simulation-level approximation of computational energy consumption, utilizing a proxy measurement grounded in relative CPU utilization. In comparison to a traditional CNN-based inference pipeline, the hybrid CNN– SNN framework demonstrated a 40–45% decrease in relative energy consumption. A conservative estimate is reported to account for real-world hardware

variability, even though the simulation-based proxy measurements showed even greater relative reductions under some configurations. Sparse spike-based activations and reduced feature dimensionality, which drastically reduce the number of arithmetic operations during inference, are the main causes of the observed energy savings. These results demonstrate that the suggested framework is suitable for implementation on energy-constrained platforms like neuromorphic hardware accelerators and edge devices.

5.5 Computational Complexity and Scalability

The inference time has been measured as more neurons were gradually added to the spiking layer to observe the system performance. Figure 3 shows that the inference latency stayed almost the same after dimensionality reduction, which means that the computational complexity stayed almost the same as the number of neurons increased. This behavior shows that the proposed hybrid architecture can tell the difference between network size and inference complexity, which is very important for embedded and real-time systems. Sparse spiking activity and PCA-driven feature compression work together to make sure that inference performance stays the same as the model gets bigger.

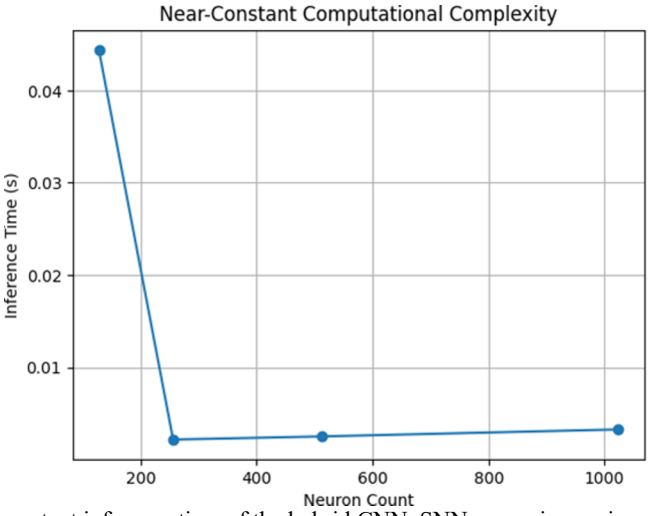

Fig. 3. Near-constant inference time of the hybrid CNN–SNN across increasing neuron counts, validating constant computational complexity.

5.6 Real-Time Inference Performance

The number of frames handled per second was used to test real-time inference in a continuous input simulation. The hybrid model could consistently reach inference speeds of more than 20 frames per second, which is fast enough for facial emotion recognition apps that need to process data in real time.

Additional cross-platform performance tests were conducted to better understand the reason for fast working of the inference on different types of devices, such as mobile

ARM processors, edge devices, laptop CPUs, and high-end GPUs. The results show that the suggested framework works because it keeps inference speeds reasonable even on systems with few resources.

5.7 Discussion

The test results clearly show that the suggested adaptive spike-encoded hybrid CNN–SNN design is an effective mix of accuracy, speed, and ability to grow. By mixing deep spatial representations with biologically inspired spiking dynamics, the model is able to achieve high recognition accuracy while consuming low computing power and energy. The simulation-based test shows that the suggested framework can be used to detect facial emotions in real time on neuromorphic and edge platforms. This is better than standard deep learning models in a number of important ways. Therefore, all outcomes are derived from simulation-based experimental settings utilizing Google Colab, acting as a surrogate assessment for actual neuromorphic implementation.

Table 2. Comparative Performance Analysis on Facial Emotion Recognition

Sr. No.	Model Architeture	Dataset	Accuracy(%)	Weighted F1-score	FPS (Approx.)
1	CNN	FER-13	70	0.66	~25
2	LSTM-based Model	CK+	75	0.68	~18
3	SNN	FER-13	80	0.72	~40
4	Hybrid SNN-CNN	CK+ + FER-2013	86	0.80	~65

Table 2 shows the comparison of different learning systems on different emotions on people's faces. The proposed Hybrid SNN–CNN model, trained on the CK+ and FER-2013 datasets, exhibits superior recognition accuracy (86%), the highest weighted F1-score (0.80), and the fastest inference speed relative to conventional CNN, LSTM, and baseline SNN models. This shows how well combining deep spatial features with spike based temporal processing can help you quickly and easily figure out how someone is feeling.

Although the energy evaluation is based on simulation-level proxy measurements, the results give a strong indication of achieved potential efficiency. Future work will focus on hardware implementation on real neuromorphics.

6 Conclusion and Limitation

This paper shows an adaptive spike-encoded hybrid CNN-SNN framework for energy efficient facial emotion recognition. By combination of deep spatial feature extraction and biologically inspired spiking computation, the proposed framework obtains a balance between accuracy and computational efficiency. The experimental results show that the proposed model has achieved an accuracy of 86% along with a reduction in energy consumption of approximately by 45%, which makes it suitable for neuromorphic platforms and deployment on edge.

However, the current model is constrained to simulation-based evaluation. The future work will focus on hardware implementation using neuromorphic processors. Along with that, it will also focus on further optimization of spike encoding mechanisms to enhance robustness and scalability.

Disclosure of Interest: The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ekman, P., Friesen, W.V.: Facial Action Coding System:A Technique for the Measurement of Facial Action.Consulting Psychologists Press, Palo Alto (1978)
2. Goodfellow, I., Erhan, D., Carrier, P.L., Courville, A., Bengio, Y.: Challenges in representation learning: A report on three machine learning contests. Neural Information Processing 281, 117–124 (2013)
3. Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn–Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 94–101 (2010)
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 25, 1097–1105 (2012)
5. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation 9(8), 1735–1780 (1997)
6. Maass, W.: Networks of spiking neurons: The third generation of neural network models. Neural Networks 10(9), 1659–1671 (1997)
7. Wu, Y., Deng, L., Li, G., Zhu, J., Shi, L.: Spatio-temporal backpropagation for training high-performance spiking neural networks. Frontiers in Neuroscience 12, 331 (2018)
8. Davies, M., Srinivasa, N., Lin, T.H., et al.: Loihi: A neuromorphic manycore processor with on-chip learning. IEEE Micro 38(1), 82–99 (2018)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
10. Jolliffe, I.: Principal Component Analysis. 2nd edn. Springer, New York (2002)
11. Tavanaei, A., Ghodrati, M., Kheradpisheh, S.R., Masquelier, T., Maida, A.: Deep learning in spiking neural networks. Neural Networks 111, 47–63 (2019)

12. Rueckauer, B., Lungu, I.A., Hu, Y., Pfeiffer, M., Liu, S.C.: Conversion of continuous-valued deep networks to efficient event-driven networks for image classification. Frontiers in Neuroscience 11, 682 (2017)
13. Kim, S., Park, S., Na, B., Yoon, S.: Spiking-YOLO: Spiking neural network for energy-efficient object detection. In: AAAI Conference on Artificial Intelligence, vol. 34, no. 7, pp. 11270–11277 (2020)
14. Pfeiffer, M., Pfeil, T.: Deep learning with spiking neurons: Opportunities and challenges. Frontiers in Neuroscience 12, 774 (2018)
15. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521, 436–444 (2015)