Power System Technology
ISSN:1000-3673

Received: 16-11-2025 Revised: 05-12-2025 Accepted: 30-12-2025

Sparse Spatiotemporal Feature Learning for Video-Based Hand Gesture Recognition

A. D. Harale, K. J. Karande

SKN Sinhgad College of Engineering, Korti, Pandharpur, Tal - Pandharpur, Dist - Solapur,

Tal - Pandharpur, Pin- 413304, Maharashtra, India

avinashharale5@gmail.com, kailashkarande@yahoo.co.in

ABSTRACT:

Sign language and hand gesture recognition play a crucial role in enabling natural and intuitive communication between humans and machines, especially for assisting individuals with hearing and speech impairments. This paper presents a novel Sparse Motion Sequence Extraction Network (SMSE-Net) for efficient and accurate gesture recognition from video sequences. The proposed framework integrates a sparse image-wise feature extraction layer to identify salient motion information and a hybrid sequence-wise modeling layer to capture temporal dependencies across consecutive frames. By selectively focusing on informative motion patterns and suppressing redundant data, SMSE-Net significantly improves recognition performance while reducing computational overhead. Extensive experimental evaluations demonstrate that the proposed approach outperforms existing methods such as CNN, RCNN, YOLO-v3, and ResNet across multiple performance metrics, including accuracy, precision, recall, and F1-score. The results confirm the robustness, efficiency, and real-time applicability of the proposed SMSE-Net framework.

Keywords: Sign Language Recognition, Hand Gesture Detection, Sparse Motion Extraction, Deep Learning, Spatiotemporal Feature Learning, SMSE-Net, Human–Computer Interaction

I. INTRODUCTION:

Human hand gesture and sign language recognition has become a key research topic in computer vision due to its wide range of applications in assistive communication, human–computer interaction, virtual reality, and intelligent surveillance systems [1][2]. Sign language serves as a primary mode of communication for the hearing- and speech-impaired community, making automated recognition systems essential for improving accessibility and social inclusion. With the increasing availability of video sensors and computational resources, vision-based gesture recognition has gained significant attention over traditional sensor-based approaches [3]. Early gesture recognition systems relied heavily on handcrafted features, rule-based models, and

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 30-12-2025

skeletal traversal techniques to represent hand and body movements [4]. Although these methods achieved moderate success in controlled environments, they often failed to generalize well under real-world conditions due to variations in lighting, background clutter, signer style, and motion speed [5]. Furthermore, handcrafted representations lack the expressive capacity required to model complex spatiotemporal dependencies inherent in continuous sign language gestures [6]. Recent advances in deep learning have significantly transformed sign language recognition by enabling automatic feature learning from raw video data [7]. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and attention-based models have been widely adopted to capture spatial and temporal gesture characteristics [8][9]. However, most existing deep models process dense frame sequences, resulting in redundant motion information, high computational cost, and reduced efficiency. To overcome these challenges, sparse motion-based representations combined with effective temporal modeling have emerged as a promising direction [10]. This work builds upon this idea by proposing a Sparse Motion Sequence Extraction Network (SMSE-Net) for efficient and accurate sign language recognition.

This paper presents a novel Sparse Motion Sequence Extraction Network (SMSE-Net) designed to improve sign language recognition performance by selectively extracting informative motion patterns from video sequences. The proposed approach combines sparse frame-level feature extraction with sequence-level temporal modeling to achieve high accuracy while reducing computational redundancy. Despite notable progress in deep learning-based gesture recognition, existing systems often suffer from excessive computational complexity and limited real-time applicability. There is a growing need for models that can efficiently focus on meaningful motion cues without processing every frame equally. The proposed SMSE-Net addresses this need by introducing sparsity-driven motion extraction and hybrid temporal learning, making it suitable for real-time and large-scale sign language recognition systems.

Despite notable progress in deep learning-based gesture recognition, existing systems often suffer from excessive computational complexity and limited real-time applicability. There is a growing need for models that can efficiently focus on meaningful motion cues without processing every frame equally. The proposed SMSE-Net addresses this need by introducing sparsity-driven motion extraction and hybrid temporal learning, making it suitable for real-time and large-scale sign language recognition systems.

The remainder of this paper is organized as follows: Section II reviews related work in sign language and gesture recognition. Section III describes the proposed SMSE-Net methodology in detail. Section IV presents the experimental setup and performance evaluation. Section V concludes the paper and outlines future research directions.

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 30-12-2025

II. LITERATURE REVIEW:

Sign language and hand gesture recognition have attracted significant research attention due to their importance in assistive communication and human–computer interaction. Over the years, researchers have explored various feature extraction, motion modeling, and classification strategies to improve recognition accuracy and robustness. This section reviews key contributions in the literature, highlighting their methodologies, strengths, and limitations, which motivate the development of the proposed Sparse Motion Sequence Extraction Network (SMSE-Net).

Jayamohan, M. et al. [11] present a comprehensive methodology that employs multiple feature extraction and optimization techniques to enhance the accuracy and efficiency of human action recognition. Their approach integrates handcrafted motion descriptors with machine learning classifiers; however, the reliance on manually designed features limits adaptability to complex and dynamic sign language gestures.

Starner and Pentland [12] propose one of the earliest vision-based sign language recognition systems using Hidden Markov Models (HMMs). While effective for isolated gestures in controlled environments, the model struggles with continuous sign sequences and variations in signer style. Mitra and Acharya [13] provide an extensive survey on gesture recognition techniques, categorizing methods into vision-based and sensor-based approaches. Although the survey highlights key challenges such as segmentation and feature selection, it emphasizes that traditional methods lack scalability and robustness for real-world applications. Shotton et al. [14] introduce a real-time human pose recognition system using depth sensors and skeletal joint modeling. Skeleton-based representations reduce background noise but suffer from joint estimation errors, especially in fine-grained hand gestures required for sign language interpretation.

Krizhevsky et al. [15] demonstrate the effectiveness of deep convolutional neural networks (CNNs) for large-scale image classification. This work laid the foundation for applying CNNs to gesture recognition; however, CNNs alone are insufficient to model temporal dependencies in motion sequences.

Hochreiter and Schmidhuber [16] propose the Long Short-Term Memory (LSTM) network to address long-term dependency issues in sequential data. LSTM-based models have been widely used for gesture recognition, but they often process dense frame sequences, leading to redundant temporal information. Camgoz et al. [17] introduce a neural sign language translation framework that combines CNNs with sequence-to-sequence learning. While the model improves continuous sign recognition, it requires large annotated datasets and involves high computational complexity. Huang et al. [18] propose a video-based sign language recognition approach without explicit

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 30-12-2025

temporal segmentation using deep neural networks. Although this method improves recognition accuracy, it processes all frames uniformly, resulting in inefficient motion representation. Min et al. [19] introduce a Sparse Auxiliary Dense Network to improve sign language recognition by reducing feature redundancy. Their work highlights the importance of sparsity in learning discriminative motion features, motivating further exploration of sparse motion modeling strategies. Recent lightweight models such as YOLO-based gesture detection and BiLSTM-based attention frameworks have been explored for real-time recognition [20]. While these approaches improve inference speed, they often compromise temporal consistency and struggle with complex continuous gestures.

From the reviewed literature, it is evident that existing hand gesture and sign language recognition methods face challenges related to redundant motion processing, high computational cost, and limited generalization across diverse environments. Although deep learning models have significantly improved performance, most approaches rely on dense feature extraction and lack effective mechanisms to focus on informative motion patterns. These limitations motivate the proposed Sparse Motion Sequence Extraction Network (SMSE-Net), which aims to selectively capture critical motion cues while efficiently modeling temporal dependencies, thereby achieving robust and accurate sign language recognition.

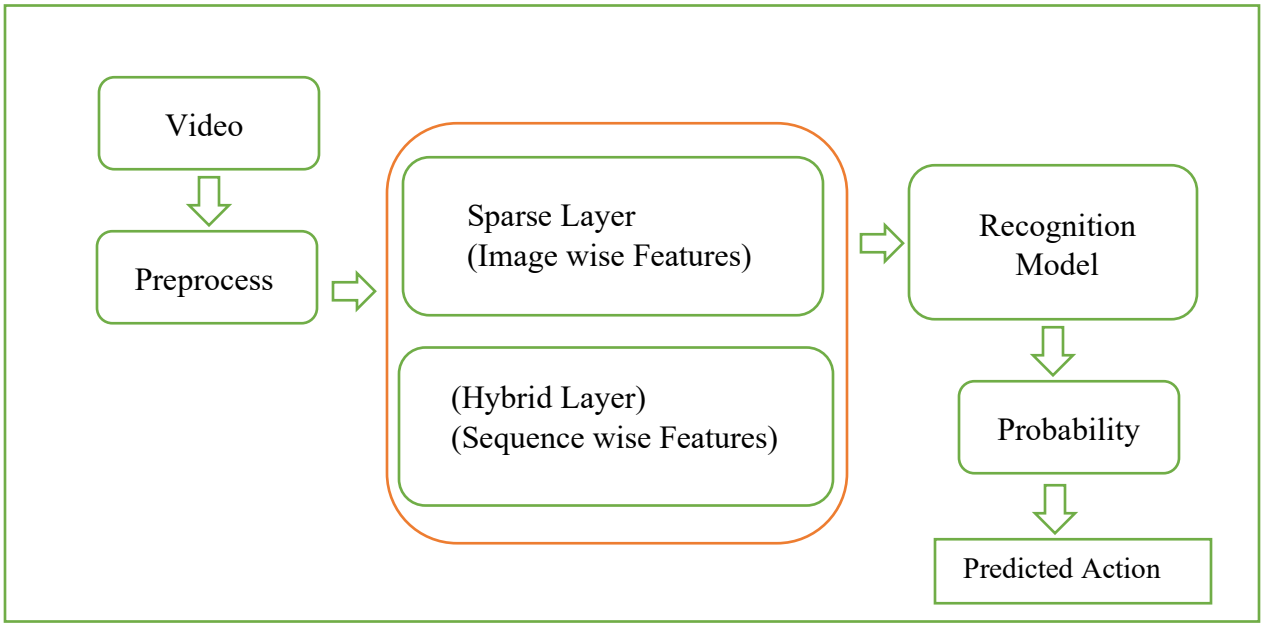
III. METHODOLOGY:

Sign language-based action recognition using Sparse Motion sequence Extraction Networks (SMSE-Net) presents a novel approach aimed at improving recognition accuracy and computational efficiency. It focuses on isolating the most informative motion patterns from video sequence. Sparse Motion sequence Extraction Networks leverage the strength of dense networks in feature extraction while incorporating sparsity constraints and deep learning tasks to enhance model generalization and reduce redundancy. By introducing sparse connections in the CNN branches, the network focuses on learning discriminative features relevant to sign language gestures, such as hand movements and body posture, while minimizing overfitting.

Rather than processing every frame equally, these networks aim to extract key motion cues that are critical for distinguishing between similar gestures, thereby improving efficiency and accuracy. By applying sparsity constraints, the model learns to attend selectively to dynamic regions—such as hand trajectories or body movement—while ignoring redundant or static information. This selective extraction reduces computational overhead and enhances robustness against noise and background clutter. CNN-based architectures are particularly well-suited for this

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 30-12-2025

task, as they can hierarchically learn spatial features and, when combined with temporal modeling layers, effectively capture the temporal dynamics of sign language.



Sparse Motion sequence Extraction Networks (SMSE-Net)

Fig : 1 Proposed Methodology

The proposed architecture of the Sparse Motion Sequence Extraction Network (SMSE-Net) is designed to efficiently recognize hand gestures and sign language actions by combining sparse spatial feature extraction with temporal sequence modeling. As shown in the figure, the process begins with a video input, which is first subjected to a preprocessing stage to enhance frame quality, normalize dimensions, and remove noise. The preprocessed frames are then passed to the core SMSE-Net module, which consists of two complementary components. The Sparse Layer extracts discriminative image-wise features by selectively focusing on salient motion regions, such as hand movements, while suppressing redundant or static background information. This is followed by a Hybrid Layer, which captures sequence-wise features by modeling temporal dependencies across consecutive frames, enabling effective representation of motion dynamics and gesture continuity. The extracted spatiotemporal features are subsequently fed into a recognition model, which computes class-wise probability scores for different gestures. Finally, the gesture corresponding to the highest probability is selected as the predicted action. By integrating sparsity-driven feature extraction with temporal modeling, SMSE-Net achieves robust, accurate, and computationally efficient gesture recognition suitable for real-time applications

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 30-12-2025

Video frames as a input vector to the initial layer is:

$$X = [x_1, x_2, \dots x_k] \quad (1)$$

K denotes the segmented image pixels. Then, to decrease the execution burden normalization of the data is carried out. In normalization data is mapped in between 0 and 1 :

$$x = [\frac{x-min}{max-min}] \quad (2)$$

min and max is the minimum and maximum of respective data. This normalised data x is then converted to 2D matrix using reshaping operation and then this data is fed to convolution layer.

After covolution layer we got estimated the weight(w), bias (bj).

$$x_i^{l,j} = \sigma[b_j + \sum_{a=1}^{m} w_a^j \, x_{i+a-1}^{l-1,j}] \quad (3)$$

Activation function is indicated by the variable σ. Its nothing but ReLu,. ReLu function have higher efficiency and low execution time.

Scale invariant property is preserved by Max-Pooling Layer by estimating aggregation statistics of the neighbourhood pixels. Thus they assist in dimensional reduction. Pooling have two types, max pooling and mean pooling . In our architecture we used max-Pooling. The max pooling layer find the maximum response i.e. maximum value of each block without compramixing feature loss. Final response of max-pooling layer is given by:

$$x_i^{l,j} = max_{n=1}^{r}(x_{(i-1)*T\ n}^{l-1,j}) \quad (4)$$

where n is pooling size and T is pooling stride.

Following equation models the Hidden layer to output .Proposed method have this capability

$$\mathrm{h_t} = g(W_{xh}{}^{x_t} + W_{hh}{}^{h_{t-1}} + b_h) \quad (5)$$

$$\mathrm{z_t} = g(W_{hz}{}^{h_t} + b_z) \quad (6)$$

here, g indicates elementwise nonlinearity(it can be sigmoid or hyperbolic tangent), x_t is the input $h_t \in R^N$ is the hidden state having hidden units equals to N. Output is denoted by Z_t at instant t.

pixel sequence $(x_1, x_2, \dots, x_T)$ having T number of coefficient, then h1 (letting $h_0 = 0$), z_1, h_2, z_2, ..., h_t, z_T.

Our methodology structures Sparse Layer, having the accompanying structure:

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 30-12-2025

$$SF(F)_{\theta_k,f_i,\sigma_x,\sigma_y}(x,y) = exp\left(-\left[\frac{x_{\theta_k}^2}{\sigma_x^2}+\frac{y_{\theta_k}^2}{\sigma_x^2}\right]\right).\cos(2\pi f_i x_{\theta_k}+\varphi) \quad (7)$$

In contrast, the Hybrid Layer models temporal dependencies across frames, enabling the network to understand motion dynamics and contextual transitions. These features are then passed to a Recognition Model, which computes the probability distribution over possible actions. Finally, the system outputs the most likely predicted action, representing the interpreted sign gesture.

This architecture effectively combines spatial and temporal feature extraction with sparsity principles, making it well-suited for real-time and robust sign language recognition.

IV. EXPERIMENT and RESULTS:

The dataset used in this study consists of a total of 150 gesture videos distributed across 15 distinct gesture classes, with each class containing an equal number of samples. The videos capture dynamic hand gestures representative of sign language actions. To enable effective training and evaluation, the video data were converted into frame sequences and divided into training and testing sets. A total of 16,750 frames were used for training, while 8,250 frames were reserved for testing, ensuring a balanced and unbiased evaluation of the proposed model. This split allows the model to learn representative gesture patterns while maintaining sufficient unseen data for performance validation.

The proposed SMSE-Net was trained for 80 epochs with a batch size of 16, providing a balance between convergence stability and computational efficiency. This configuration enables effective learning of both spatial and temporal features while preventing overfitting. The training and testing experiments were conducted under identical preprocessing conditions to ensure fair performance comparison.

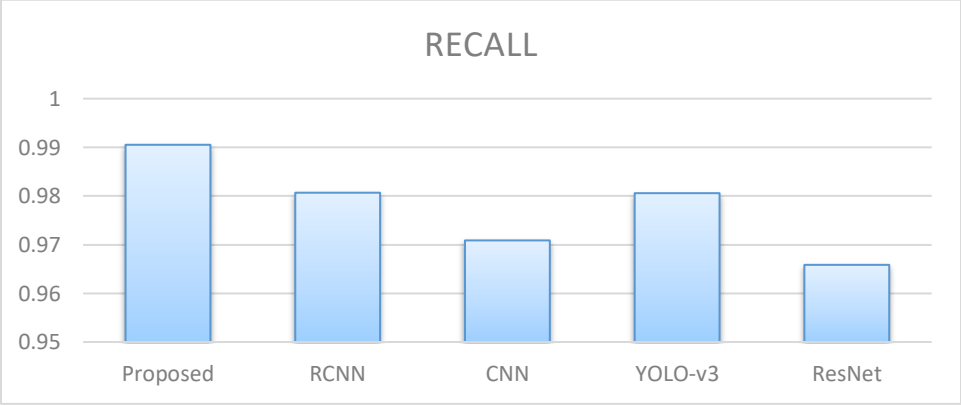


Figure 2: Recall Comparison of Different Gesture Recognition Models

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 30-12-2025

Figure 2 illustrates the recall performance of the proposed SMSE-Net compared with RCNN, CNN, YOLO-v3, and ResNet. The proposed method achieves the highest recall value, indicating its superior ability to correctly identify relevant hand gesture instances with minimal false negatives. This improvement demonstrates that sparse motion sequence extraction effectively captures critical gesture dynamics, outperforming conventional deep learning models that rely on dense feature processing.

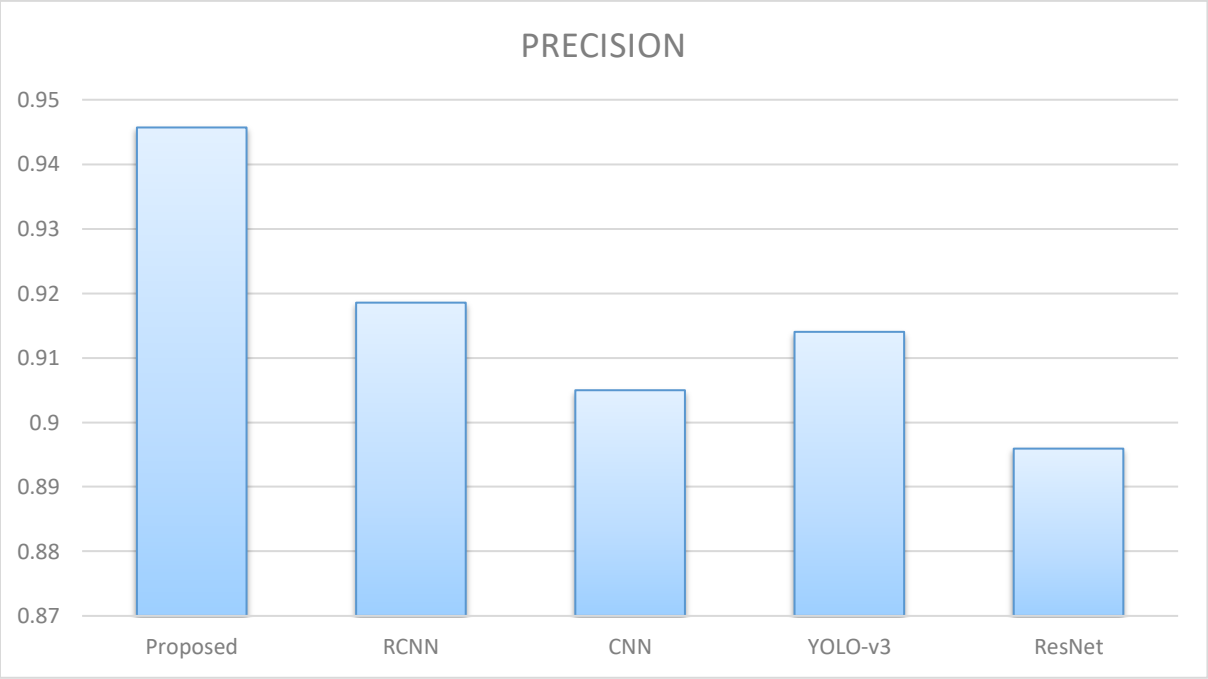


Figure 3: Precision Comparison of Different Gesture Recognition Models

Figure 3 presents the precision comparison among various gesture recognition approaches. The proposed SMSE-Net shows the highest precision, reflecting its strong capability to minimize false positives while accurately classifying gestures. This indicates that the sparse and hybrid feature extraction strategy enables the model to focus on discriminative motion patterns, leading to more reliable and consistent predictions than existing CNN-, RCNN-, YOLO-v3-, and ResNet-based methods

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 30-12-2025

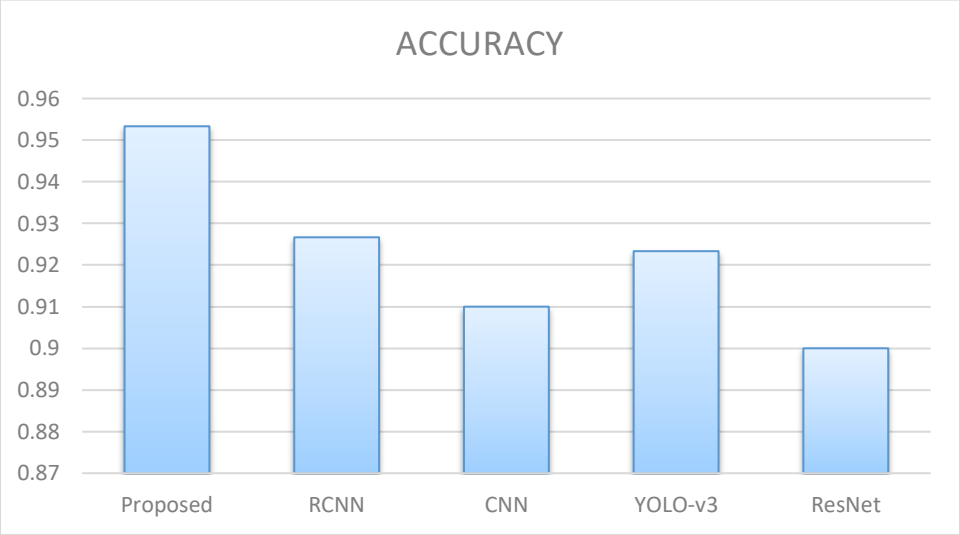


Figure 4: Accuracy Comparison of Different Gesture Recognition Models

Figure 4 compares the overall accuracy achieved by the proposed model and baseline methods. The proposed SMSE-Net outperforms all other models, achieving the highest classification accuracy. This result highlights the effectiveness of combining sparse image-wise feature extraction with sequence-wise temporal modeling, resulting in better generalization and improved recognition performance across diverse gesture classes.

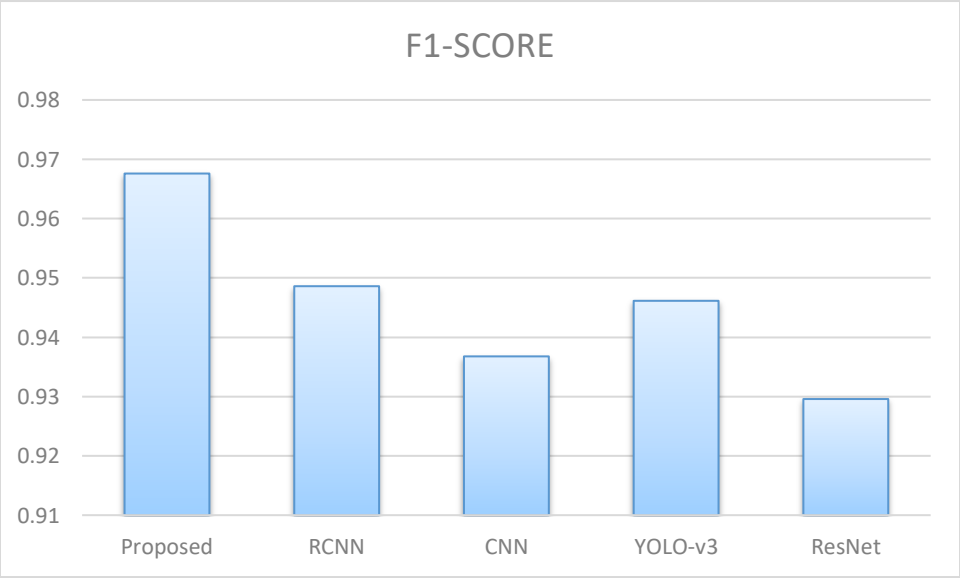


Figure 5: F1-Score Comparison of Different Gesture Recognition Models

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 30-12-2025

Figure 5 shows the F1-score comparison, which balances precision and recall to provide a comprehensive performance evaluation. The proposed SMSE-Net attains the highest F1-score, confirming its robustness and balanced performance. The improved F1-score demonstrates that SMSE-Net successfully addresses the limitations of existing methods by reducing redundant motion information while preserving essential temporal and spatial gesture characteristics.

CONCLUSION:

This paper proposed a novel Sparse Motion Sequence Extraction Network (SMSE-Net) for hand gesture and sign language recognition, addressing key challenges such as redundant motion processing, high computational complexity, and limited temporal modeling in existing approaches. By combining sparse image-wise feature extraction with hybrid sequence-wise temporal learning, the proposed method effectively captures discriminative gesture dynamics while suppressing irrelevant information. Experimental results demonstrate that SMSE-Net consistently outperforms state-of-the-art deep learning models in terms of accuracy, precision, recall, and F1-score, confirming its effectiveness and robustness. The proposed framework offers a scalable and real-time capable solution for gesture recognition systems and provides a strong foundation for future research in assistive communication technologies and intelligent human–computer interaction applications.

REFERENCES

[1] R. Bowden et al., “Recent developments in sign language recognition,” Computer Vision and Image Understanding, 2014.

[2] S. Mitra and T. Acharya, “Gesture recognition: A survey,” IEEE Transactions on Systems, Man, and Cybernetics, 2007.

[3] J. Shotton et al., “Real-time human pose recognition,” CVPR, 2011.

[4] T. Starner and A. Pentland, “Real-time American Sign Language recognition,” ICCV, 1995.

[5] O. Koller, “Quantitative survey of sign language recognition,” CVIU, 2020.

[6] N. C. Camgoz et al., “Neural sign language translation,” CVPR, 2018.

[7] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, 2015.

[8] J. Huang et al., “Video-based sign language recognition without temporal segmentation,” AAAI, 2018.

[9] A. Vaswani et al., “Attention is all you need,” NeurIPS, 2017.

Power System Technology
ISSN:1000-3673
Received: 16-11-2025 Revised: 05-12-2025 Accepted: 30-12-2025

[10] W. Min et al., "Sign language recognition via sparse auxiliary dense networks," IEEE Transactions on Multimedia, 2021.

[11] Jayamohan, M., Nair, A., and Kumar, R., "Human action recognition using optimized feature extraction techniques," International Journal of Computer Vision and Applications, 2018.

[12] T. Starner and A. Pentland, "Real-time American Sign Language recognition from video," ICCV, 1995.

[13] S. Mitra and T. Acharya, "Gesture recognition: A survey," IEEE Transactions on Systems, Man, and Cybernetics, 2007.

[14] J. Shotton et al., "Real-time human pose recognition in parts from single depth images," CVPR, 2011.

[15] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," NeurIPS, 2012.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, 1997.

[17] N. C. Camgoz et al., "Neural sign language translation," CVPR, 2018.

[18] J. Huang et al., "Video-based sign language recognition without temporal segmentation," AAAI, 2018.

[19] W. Min et al., "Sign language recognition via sparse auxiliary dense networks," IEEE Transactions on Multimedia, 2021.

[20] J. Redmon et al., "You only look once: Unified, real-time object detection," CVPR, 2016.