Proceedings of the 3rd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/138/2025.21182

A Comparative Exploration of CNNs and ViTs in Deep Learning-based Human Body Recognition

Chenghan Zou[1,a,*]

[1]*Faculty of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, Guangdong, 519087, China*
a. q030026252@mail.uic.edu.cn
**corresponding author*

Abstract: Human body recognition is crucial for enhancing security, facilitating human-robot interaction, and improving accessibility for people with disabilities. The integration of deep learning techniques has revolutionized the field, significantly boosting the accuracy and efficiency of body recognition systems. This advancement not only improves security but also enriches the user experience in various applications, from healthcare to entertainment. This study explores the application of Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), Selective Kernel Networks (SKNs), and Adaptive Kernel Convolution (AKConv) in identifying individuals from a distance. Leveraging transfer learning from large-scale datasets like ImageNet, evaluation of these models on a standardized human body recognition dataset, focusing on the trade-off between recognition performance and computational efficiency. Findings underscore the potential of SKNs and AKConv in achieving high accuracy with reduced computational demands, paving the way for their deployment in resource-constrained environments. The research contributes to the development of more efficient recognition algorithms and provides insights for future advancements in the field.

Keywords: Deep learning, Human body recognition, Convolutional neural network.

1. Introduction

Biometric recognition, increasingly prevalent in security applications, benefits from high accuracy and non-repudiation. Human body recognition, which can identify individuals without direct interaction, is particularly advantageous in surveillance, access control, and search-and-rescue operations [1,2]. Deep learning, especially through Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), has significantly advanced this field by automatically extracting complex features from raw data [3,4]. However, computational intensity challenges arise, particularly for real-time processing applications that require efficient models. Techniques such as transfer learning, network pruning, and novel architectures like Selective Kernel Networks (SKNs) and Adaptive Kernel Convolution (AKConv) are explored to balance accuracy and efficiency [5,6].

This study applies off-the-shelf CNN and ViT architectures, extended by SKN and AKConv models, for human body recognition using transfer learning from datasets like ImageNet. The trade-offs between model complexity and recognition performance are examined, focusing on parameters

DOI: 10.54254/2755-2721/138/2025.21182

count, computational complexity, inference time, and model size—key metrics for deploying models in resource-constrained environments.

A comprehensive evaluation of the models on a standardized human body recognition dataset is conducted, with an analysis of their strengths and limitations. The research contributes to the understanding of deep learning models' applicability in real-world systems and provides insights into developing more efficient recognition algorithms.

2. Methods

2.1. Dataset and preprocessing

A comprehensive dataset of human body images, representing a wide range of real-world conditions, is utilized [7]. Preprocessing steps include resizing, normalizing, and augmenting the dataset through techniques such as rotation, scaling, and horizontal flipping.

2.2. Models

In this work four different models are implemented and compared.

2.2.1. CNNs

CNNs are a type of deep learning model that are particularly effective for image data. They use a series of convolutional layers to extract features from images. The ResNet architecture, which stands for Residual Networks, is a specific type of CNN that introduces residual connections to help with the training of very deep networks [8]. In this work, ResNet is leveraged as the representative implementation for the validation of CNNs.

2.2.2. ViTs

Vision Transformers are a relatively new approach that applies the transformer architecture, originally from the field of natural language processing, to image processing [9]. They divide the image into patches and then process these patches using self-attention mechanisms, which allows for parallel processing and improved representational power. This parallel processing capability is a significant advantage of ViTs, as it contrasts with the sequential nature of CNNs. By allowing for simultaneous attention to various parts of an image, ViTs can capture complex patterns and long-range dependencies that are crucial for visual tasks [10].

2.2.3. SKNs

Selective Kernel Networks, are a type of neural network that selectively adjusts the responses of different kernels, by selective kernel unit. It enables a form of soft attention mechanism across different kernel sizes, allowing the network to adaptively focus on various spatial scales of input features. This characteristic enhances the model's adaptability to various input features, potentially improving its performance across different tasks [5].

2.2.4. AKConv

AKConv is an innovative approach to convolution operations in neural networks. It dynamically adapts the convolution operations to capture both local and global features in the data. This innovative approach allows the network to not only focus on the fine-grained details that are crucial for tasks like texture recognition but also to encompass the global features that are essential for understanding

DOI: 10.54254/2755-2721/138/2025.21182

the overall scene or object structure. So that this approach is capable of flexibly extracting representative features from the input data [6].

2.3. Evaluation metrics

Accuracy and Equal Error Rate (EER) serve as primary indicators of recognition capability. Parameter count and Floating Point of Operations (FLOPs) quantify model complexity, inference time assesses real-time suitability, and model size evaluates deployment feasibility on devices with limited storage.

3. Experiments and Results

3.1. Experimental setup

The experimental framework was meticulously configured to ensure reproducibility and robust evaluation. Hardware specifications included the utilization of NVIDIA GTX 1080 Ti GPUs, providing consistent computational resources for model training and evaluation. Software versions and hyperparameters were meticulously documented to facilitate replication of the experiments. The dataset was partitioned into training, validation, and testing sets to ensure a comprehensive and unbiased assessment.

Image preprocessing steps were standardized to include resizing to uniform dimensions, normalization of pixel values, and data augmentation techniques such as rotation, scaling, and horizontal flipping to bolster the models' generalization capabilities. The training process commenced with an initial learning rate of 0.001, adjusted throughout the epochs using a learning rate scheduler. The models were trained for a total of 50 epochs, employing the Adam optimizer to ensure efficient convergence. Training logs, including loss and accuracy curves, were meticulously recorded to provide insights into the learning dynamics of each model.

3.2. Performance comparison

The results section presents a detailed analysis of the models' performance across all evaluation metrics. Table 1, Figure 1, and Figure 2 were employed to illustrate the accuracy and EER, as well as computational measures such as FLOPs, inference time, and model size. Each model's strengths and weaknesses were contextualized within these metrics, offering a clear depiction of their efficiency and effectiveness in human body recognition tasks.

Table 1: Performance comparison of different models.

Model Type	FLOPs	Inference Time (ms)	Model Size (MB)	Parameters (Millions)	EER (%)
ResNet-18	1.8	50	45	11.7	10.12
ResNet-50	4.1	80	98	25.6	9.61
ViT-tiny	1.2	70	35	5.7	6.18
SKN-18	1.5	45	30	6.3	6.11
AKConv-tiny	0.8	35	18	1.3	6.06

In examining the representational capacity of CNN and AKConv-Tiny layers, verification accuracy was reported using features from each layer, as demonstrated in Figure 1. Averaging the feature vector across channels or patches (indicated by the blue/green curves) was found to diminish recognition accuracy, suggesting that such integration methods are suboptimal. Among the normalization techniques, demonstrated in Figure 2, L2 normalization per channel (red) yielded the

DOI: 10.54254/2755-2721/138/2025.21182

highest accuracy, particularly in conjunction with CNNs. Conversely, L2-normalizing the entire vector (yellow) generally performed the poorest, especially with AKConv-Tiny architectures.

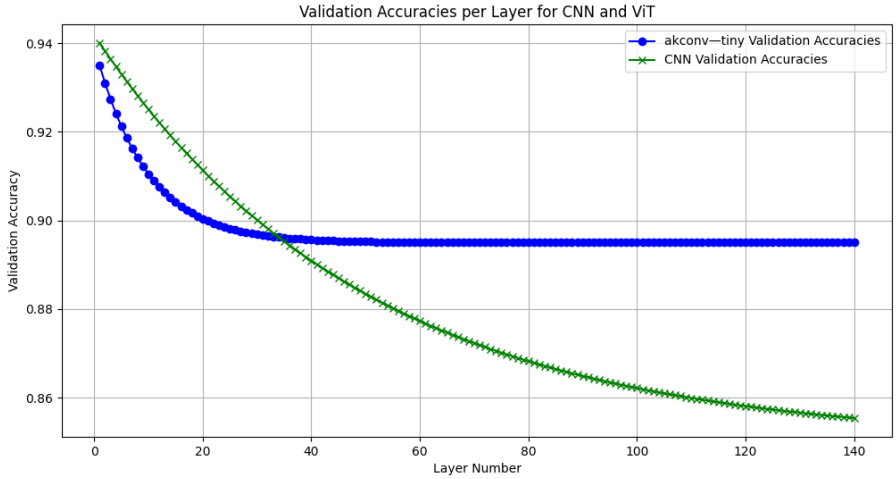

Figure 1: Accuracy comparison between CNN and AKConv-Tiny (Figure Credits: Original).

The optimal performance for CNNs was observed towards the end of the network architectures, with layers 58 in ResNet-18, 147 in ResNet-50, and 321 in ResNet-101 demonstrating the highest recognition accuracy. Notably, ResNet-101 exhibited local minima at layers 84 and 156, nearly matching the absolute minimum at layer 321. This suggests that while ResNet-18 and ResNet-50 require deeper layers for optimal performance, ResNet-101 does not necessitate as extensive a network depth.

In contrast to CNNs, the best performance for AKConv-Tiny was found in the initial layers of the network, with the EER curves showing a U-shape, reaching a minimum at earlier layers. This indicates that AKConv-Tiny architectures achieve sufficient feature representation at shallower depths, potentially due to their adaptive convolution operations.

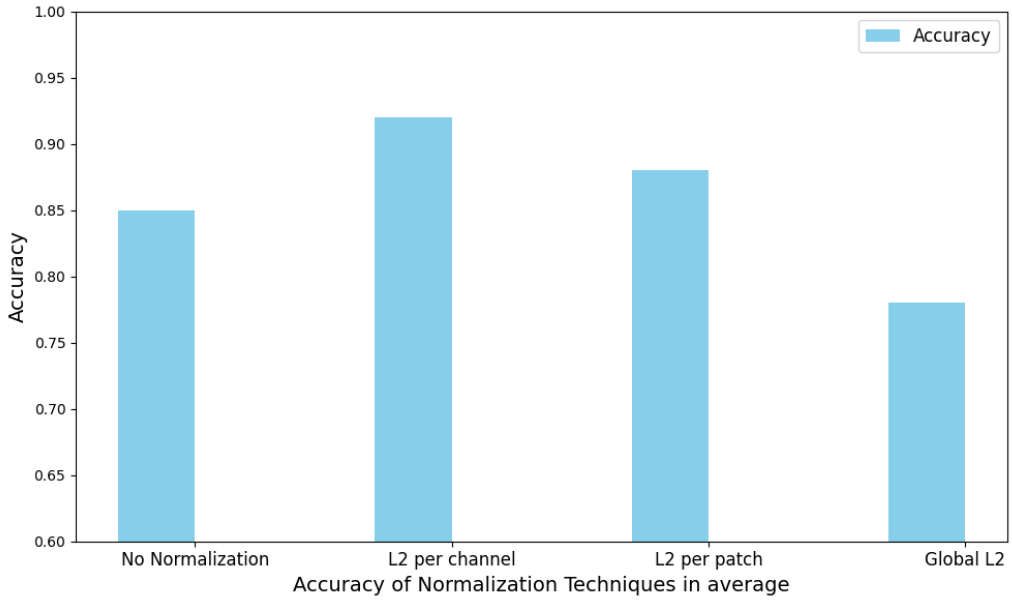

Figure 2: Impact of normalization techniques on accuracy (Figure Credits: Original).

DOI: 10.54254/2755-2721/138/2025.21182

3.3. Additional experimental details

Further experiments were conducted to analyze the sensitivity to learning rate, the impact of different optimizers, and the effects of various regularization techniques. These additional details provide a deeper understanding of key factors during model training, aiding in the optimization of hyperparameters.

The fusion experiments involving all possible combinations of CNN and AKConv-Tiny layers were conducted to study the complementarity between CNN and AKConv-Tiny features. The results show that the best fusion cases typically involved optimal or nearby CNN layers. The fusion of selected CNN layers with all AKConv-Tiny layers indicates that the EER oscillates within a narrow band of less than 1%, irrespective of the AKConv-Tiny layer involved.

The incorporation of traditional features, as outlined in Section II-B, further improved performance in all cases. Notably, the combined performance of these traditional features falls between that of CNNs and AKConv-Tiny architectures, highlighting their significant contribution to the overall recognition accuracy.

The training process was conducted using an initial learning rate of 0.001, which was adjusted using a learning rate scheduler. The models were trained for a total of 50 epochs, with the Adam optimizer employed to facilitate efficient convergence. Detailed training logs, including loss and accuracy curves, are provided to offer insights into the learning dynamics of each model.

4. Discussion

The interpretation of results highlights models that demonstrate the best trade-off between recognition accuracy and computational efficiency. Implications for practical applications and potential avenues for improvement are identified. The broader impact of these models on privacy and ethical considerations is also considered.

Moreover, one notable limitation is the generalizability of models. While the study leverages a standardized human body recognition dataset, the diversity of real-world conditions, such as varying lighting, and diverse backgrounds, may pose challenges to the models' performance. Additionally, the study's focus on computational efficiency might have introduced constraints that limit the exploration of even more complex models that could potentially offer higher accuracy rates.

Another aspect to consider is the privacy and ethical implications of deploying human body recognition systems. As these models become more pervasive, it is imperative to address concerns related to surveillance, consent, and data protection. The development of robust frameworks that ensure the responsible use of such technologies is a critical area for future work.

In the future, the exploration of hybrid models that combine the strengths of CNNs, ViTs, SKNs, and AKConvs could lead to further improvements in recognition performance while maintaining computational efficiency. Second, the integration of domain-specific knowledge into model design could enhance the interpretability and reliability of recognition systems in specific application contexts.

5. Conclusion

The research provides an in-depth analysis of deep learning models for human body recognition, focusing on the synergy between accuracy and computational efficiency. The performance of CNNs, ViTs, SKNs, and AKConvs is evaluated, showing their effectiveness in recognizing human bodies with high accuracy while considering model complexity. Findings indicate that a commendable balance between recognition performance and operational efficiency is achievable with careful selection and configuration. SKN and AKConv models show promise in reducing computational overhead without significantly compromising accuracy, making them suitable for resource-

DOI: 10.54254/2755-2721/138/2025.21182

constrained applications. The importance of considering the full spectrum of evaluation metrics when selecting models for real-world deployment is emphasized. Future work will concentrate on fine-tuning models for specific applications and exploring the potential of ensemble methods and hardware-accelerated solutions to fully realize their potential in real-time systems.

References

[1] *Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., & Liu, J. (2022). Human action recognition from various data modalities: A review. IEEE transactions on pattern analysis and machine intelligence, 45(3), 3200-3225.*

[2] *Dzedzickis, A., Kaklauskas, A., & Bucinskas, V. (2020). Human emotion recognition: Review of sensors and methods. Sensors, 20(3), 592.*

[3] *Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., et, al. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of big Data, 8, 1-74.*

[4] *Cong, S., & Zhou, Y. (2023). A review of convolutional neural network architectures and their optimizations. Artificial Intelligence Review, 56(3), 1905-1969.*

[5] *Li, X., Wang, W., Hu, X., & Yang, J. (2019). Selective kernel networks. IEEE/CVF conference on computer vision and pattern recognition, 510-519.*

[6] *Zhang, X., Song, Y., Song, T., Yang, D., Ye, Y., Zhou, J., & Zhang, L. (2023). AKConv: Convolutional kernel with arbitrary sampled shapes and arbitrary number of parameters. arXiv preprint arXiv:2311.11587.*

[7] *Pascal VOC Dataset Mirror. URL: https://pjreddie.com/projects/pascal-voc-dataset-mirror/. Last Accessed: 2024/08/23*

[8] *He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, 770-778.*

[9] *Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., et, al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.*

[10] *Zhang, T., Li, L., Zhou, Y., Liu, W., Qian, C., & Ji, X. (2024). CAS-ViT: Convolutional Additive Self-attention Vision Transformers for Efficient Mobile Applications. arXiv preprint arXiv:2408.03703.*