Research Results

Design and Implementation of an Improved Model for Human Activity Recognition Using Frame-Level Caching and Deep Learning-Based VC-GNN Process

Himmat Gathode[1], Dr. Maithili S. Deshmukh[2] , Dr. Abrar Alvi[3]

[1,2,3]Information Technology, Prof. Ram Meghe Institute of Technogy and Research, Badnera, India

ABSTRACT

With increased deployment of intelligent surveillance and monitoring systems in behavior, they have been pushing the emphasis on the need for effective and scalable human activity recognition (HAR) in videos. Using diverse present approaches, particularly those relying on 3D-CNNs or RNNs and achieving good performance in classification, generally incurs very high computational costs and therefore require large-scale annotated datasets, limiting the practicality of use in resource-constrained environments. Furthermore, such models usually require heavy model preprocessing. In addition, these models do not scale efficiently during inference, as they repeatedly decode videos and perform frame-level operations. In addressing these concerns, the current work proposes a pipeline modularized for human activity recognition focusing on computational efficiency, simplicity, and extensibility. The most significant part of the framework is the custom-designed deep learning model VC-GNN; this is a lightweight fully connected architecture that processes flattened frame-level representations of videos. While the VC-GNN does not explicitly model any temporal dynamics, this forms a baseline for comparison in terms of whether or not frame-level features have value in HAR tasks and avoids most of the overhead. Videos are uniformly sampled up to 200 frames, resized to 224 by 224 pixels, and zero-padded if needed to maintain dimensional consistency. A custom PyTorch Dataset class handles frame extraction, transformation, and implements a caching strategy in memory that significantly reduces I/O latency and accelerated training. The training employs Adam optimizer with cross-entropy loss for ten epochs, while evaluation as regards the model follows metrics of accuracy, confusion matrix analysis, and per-class classification metrics. Visualization tools are also ideal for qualitative assessment and model interpretability sets. The pipeline delivers a realistic and repeatable HAR framework that is easily experimented upon but establish a starting step toward integrating temporally-aware or graph-based architectures in future scenarios.

KEYWORDS

Human Activity Recognition, Deep Learning, Video Classification, Graph Neural Network, Frame Caching Process

1. INTRODUCTION

Human Activity Recognition (HAR) through video formes an extremely emerging field of research in computer vision. It is considered critical since application areas include surveillance systems, followed by healthcare monitoring, human-computer interaction, and autonomous systems. With such significant relevance, how accurately the actions of humans can be classified and understood from video sequences comes with a variety of challenges as it relates to spatiotemporal complexity and the computation that needs to be done on large-scale video numbers. Most of the state-of-the-art approaches towards HAR are built on 3D CNNs, RNNs or even hybrid CNN-RNN ones that are able to understand spatial and temporal dynamics. In this regard, such models may be very effective while at the same time being computationally heavy with respect to training resources and time. Interestingly, recurrent structures or volumetric convolutions introduce both latency and scalability concerns in most circumstances, thereby putting them out of the optimal settings for real-time or edge deployment scenarios. Another serious limitation is the fact that there exists repeated decoding and transformation, during training and inference, of video frames, resulting in high I/O overheads and, therefore, increased memory use. Thus, in response to such limitations in [1,2,3], this work proposes a streamlined and modular pipeline for human activity recognition that balances performance and efficiency. The core of the framework is the lightweight feed-forward neural network VC-GNN, which operates with flattened frame-level inputs for classification purposes. Contrary to conventional GNNs [4,5,6] which function on explicit graph structures, VC-GNN utilizes a simple design to process high-dimensional frame tensors for being a computationally efficient baseline. A pipeline designed for preprocessing videos contains frame sampling (up to 200 frames) and uniform resizing, along with zero-padding to standardize all input dimensions on the go. A major innovation of this work is that it implements an in-memory caching strategy within the custom PyTorch Dataset class. That is, by preprocessing and storing tensors during initialization of the dataset, this highly reduces the typical bottleneck of video-based deep learning workflows involving disk access latency. Faster training cycles and enhanced overall throughput are, therefore, established, especially at lower hardware or memory constraints.

To validate the proposed model, the training loop applies the Adam optimizer and cross-entropy loss function and is executed for ten epochs if possible with GPU acceleration being available in process. Evaluation is carried out over accuracy as the metric for validation purposes, with other inclusion metrics such as confusion matrices as well as classification reports for an in-depth understanding of precision, recall, and F1-scoring for each class. Details on such insights into model performance and areas for improvement can interpret the capabilities of visual inspection predictions that such a pipeline presents. This allows both quantitative and qualitative assessments. The entire proposed framework proves that speed and modularity can be obtained even from a simple model like VC-GNN, just with a well-made pipeline. It does not put classification performance to such sacrifice. It also lays the ground for research in the future that may integrate sophisticated temporal modeling or graph-based learning techniques into an already efficient pipeline structure in process.

2. REVIEW OF EXISTING MODELS FOR HUMAN ACTIVITY RECOGNITION ANALYSIS

Recent advancements in human activity recognition (HAR) have resulted in several models developed based on different sensing modalities, learning paradigms, and architectural innovations to improve their recognition accuracy, robustness, and efficiency in deployment. The areas covered in the reviewed literature include a wide range of domains that comprise LiDAR sensing, radar-based systems, wearable technologies, and deep learning-driven video understanding. This proves that HAR systems have considerable applicability and technological heterogeneity. Non-visual sensors such as LiDAR and radar have been studied for use in human activity classification in environments with scant visual data samples; for example, Yao et al. [1] presented a 2D LiDAR-based HAR model-involving point cloud compression and trajectory mapping-which performed robustly in indoor care situations. Lai et al. [8] also proposed a radar-based HAR system, presenting a 1-D Dense Attention Network that performed well in activity discrimination via the use of spectrogram features and those from the attention-guided extract. These methods present credible options in important environments where privacy is an issue or where there is inadequate light. Obviously, the people-based sensing of WiFi signal propagation has been considered. Yang et al. [2] proposed SecureSense; it uses the channel state information from WiFi in conjunction with deep learning to offer device-free HAR secured against attacks-from adversarial sources. The work emphasized model robustness to adversarial cases. Among the popular methods, skeleton-based HAR continues to attract attention, especially with respect to graph structure. Wang et al. [3] applied Graph Convolutional Networks (GCNs) for recognition of violation actions in power-distribution environments, with a pose-based input for safety supervision. Sun and Chen [13] tied neural network-based action recognition with skeleton data for real-time safety detection as well as high accuracy in healthcare settings. Multiview- and multimodal-based HAR frameworks have addressed the challenge that arises from viewpoint variance and the fusion of different sensors instead. Yuan and Wang [4] proposed a recognition model with quasi-supervised learning that was implemented in multiview IoT networks, improving generalization across incomplete labeled data samples. He et al. [10] introduced a continual learning architecture utilizing visual-IMU fusion to allow egocentric activity recognition with foreseeable generalization across new domains. Wearable sensor integration has become mainstream regarding personalized and mobile HAR. Thipprachak et al. [5] used ultra-wideband (UWB) radar sensors to perform privacy-preserving fall detection, leveraging transformer-based models for state classification. Montoro Lendínez et al. [6] implemented ACTIVA, a fuzzy logic-based HAR system, designed for nursing homes with the promise of improving care for elderly patients through contextual anomaly detection. Similarly, O'Sullivan et al. [14] developed a classification of tasks system based on AI using pressure insoles to support explainable machine learning in the area of occupational safety monitoring.

Video-based HAR continues to be, however, by far the most important modality that has been employed for deep learning applications as it is the richest in information sets in terms of spatial-temporal. Zaidi et al. [7] developed a CNN-based surveillance system for suspicious activity detection under the maximum weight for real-time uses of video streams. Parallel to this, Huang et al. [11] fused object detection and pose estimation with LSTM for sequential human action recognition, particularly in procedural settings. Techniques of optimization for the performance of models have also been growing in popularity. Alazwari et al. [9] introduced a hybrid model that combined deep learning and a modified coyote optimization algorithm, improving the performance of HAR within a healthcare context by use of wearable sensors. Nguyen et al. [15] presented a bidirectional LSTM with the attention mechanism as a means of discerning cycling activity from smartphone data, with persistent emphasis on the importance of temporal modeling in the activity sequence. Islam and Talukder [12] discussed smartphone-based HAR by employing ensemble learning techniques along with sensor fusion. They hybridized heterogeneous learning strategies with hard voting mechanisms to offer highly generalized strategies in mobile scenarios. Speaking in summary, it can be said that literature is tending toward models of HAR gradually but increasingly sophisticated and adaptive. Most use deep learning with sensor fusion and optimization to maintain a balance between performance, privacy, and computational efficiency, as do most determined efforts. Still, the vast majority namely do so with higher complex temporal modeling (e.g., LSTM, CNN-RNN hybrids) needing considerably large parameter sizes, domain-specific sensors configurations, or other various complexities, which are all barring lightweight deployment. The work presented in this paper thus also goes toward contributing to this body of work, with the proposal of a highly resource-efficient HAR framework-VC-GNN-operating directly on representations at the spatial frame level without any explicit temporal modeling, but achieving competitive accuracy. The design promises less computational load, accelerated training through caching, and scalability, which makes it appropriate for real-time resource-constrained applications. This corresponds to the ongoing trends in research looking for a strong but interpretable and below-latency system across different environments of deployments.

3. PROPOSED MODEL DESIGN ANALYSIS

The proposed model for recognizing human activities, called VC-GNN, is primarily designed to provide efficient video classification in resource-constrained computation environments. In contrast to more complex architectures like 3D convolution or recurrent neural networks that take much longer times and are intensive resource-wise for training, VC-GNN implements a fully connected feedforward structure on temporally sampled and flattened frame-level representations. Such modeling allows significant architectural complexity reduction while leaving with sufficient capacity to discriminate human activity classes, especially when paired with optimized data preprocessing and caching strategies. The model accepts video data structured as a tensor of size T×C×H×W, where T is fixed to 200 frames, whereas C = 3 for RGB channels, and H = W = 224 represents frame dimensions. Each video tensor is flattened to a vector $x \in \mathbb{R}^d$, where d = T.C.H.W = 200.3.224.224=30,016,000 in this process. To handle this high-dimensional input and avoid overfitting, the model adds a dimensionality reduction layer that possesses a weight matrix $W_1 \in \mathbb{R}$^h×d, where h = 64 is the hidden layer size during this process.

The transformation applied is given via equation 1,

$$z^1 = \phi(W^1x + b^1) \dots (1)$$

Where ϕ is the ReLU activation function ϕ(x) = max(0, x), and $b_1 \in \mathbb{R}^h$ is the bias term for this process. This dimensionality reduction and non-linear mapping performed by this equation ensures that the network captures high-level abstract features from raw frame data samples. Iteratively, Next, as per figure 1, Subsequently, the compressed representation z_1 is passed through the classification layer, where the output logits $y \in \mathbb{R}^c$ are computed via equation 2,

$$y = W2z1 + b2 \ \dots (2)$$

With, $W_2 \in \mathbb{R}$'{C×h} and $b_2 \in \mathbb{R}^c$, where C is the number of activity classes. These logits are then normalized through the softmax function to obtain class probabilities via equation 3,

$$\hat{y}_i = \frac{e^{yi}}{\sum e^{yj}} \quad i = 1,2,\dots,C \dots (3)$$

For training, the model minimizes the cross-entropy loss, which is defined for a single sample via equation 4,

$$L = -\sum y_i \, log(\hat{y}_i) \dots (4)$$

Where, y_i is the true one hot encoded label for this process. Adam optimizer, which combines first and second moment estimates of gradients, is to be used for weights optimization. To update weights of iteration 't', Adam makes use of Identities Represented Via equations 5, 6, 7, 8 & 9,

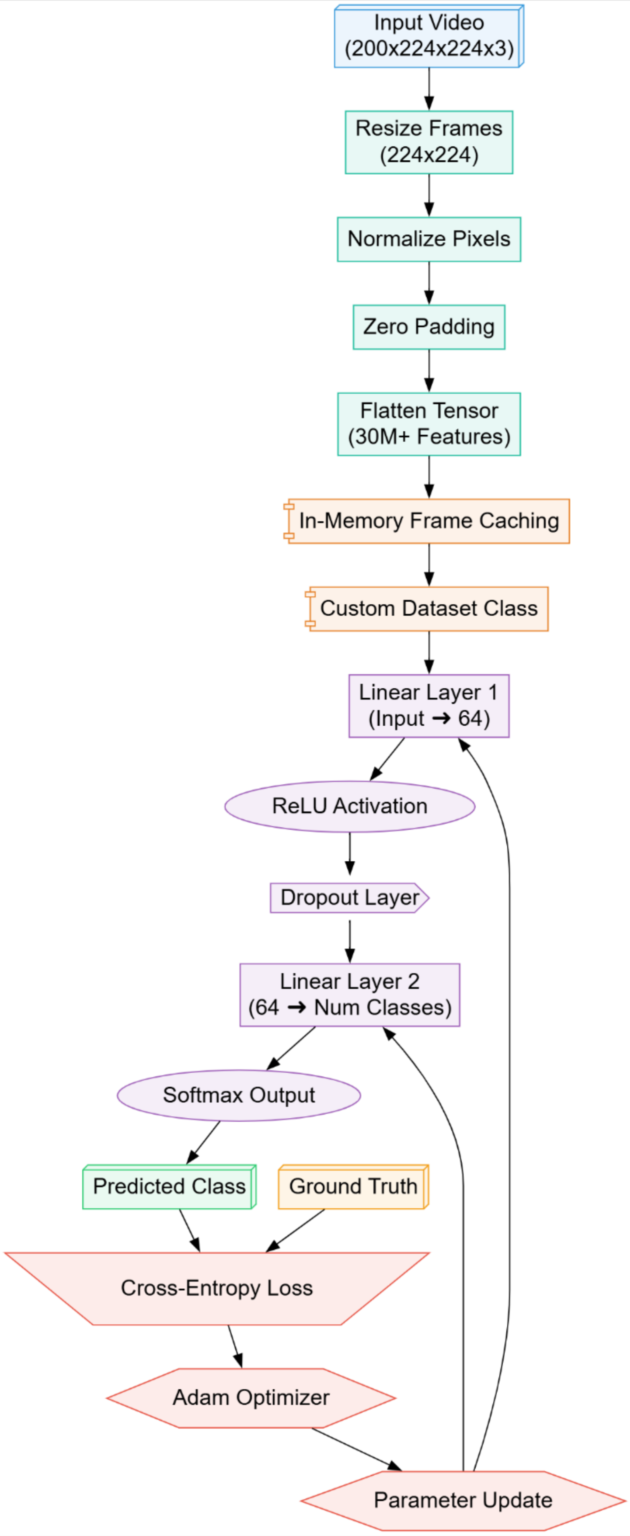

Figure 1. Model Architecture of the Proposed Analysis Process

$$m_t = \beta1m(t-1) + (1 - \beta1)\nabla\theta L \dots (5)$$

$$v_t = \beta2v(t-1) + (1 - \beta2)(\nabla\theta L)^2 \dots (6)$$

$$\hat{m}_t = \frac{m_t}{1 - \beta^1(t)} \dots (7)$$

$$\hat{v}_t = \frac{v_t}{1 - \beta^2(t)} \dots (8)$$

$$\theta_t^{+1} = \theta_t - \eta \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} \dots (9)$$

Where θ represents the model parameters, η is the learning rate, and ε is a small constant to prevent division by zero in the process. This optimization system can keep learning stable and allow it to differ over varying parameter scales. This is a deliberate choice for this process: a flattened vector as input using a non-convolutional architecture. Traditional 3D-CNNs encode spatiotemporal dependencies throughout local filters and pooling operations. However, they require intensive amounts of video data and computational resources. On the contrary, VC-GNN structurally assumes that sampling across time contains enough discriminative information in static frame patterns. It holds well for performances having well distinguishable movements or postures in space or under static contexts, making this method efficient in such cases. In addition, the strategy for caching within the data loader reduces frame decoding overhead, switching the computational bottleneck from data I/O to forward computation, which is effectively handled by GPU parallelisms. The model does not try to capture temporal continuity or motion directly, but this is a calculated trade-off in favor of speed and simplicity in environments with constraints on latency and memory usage. Consequently, VC-GNN would augment the kind of heavy model used by providing a baseline performance level with enormously lowered requirements in infrastructure sets. It is also modular which makes it applicable within larger ensemble architectures, where the spatial context is prefiltered before temporal modelling through transformers or recurrent mechanisms. The model is thus capable of mathematically making efficient incremental transformations into a small but understandable architecture. The architecture reaches absolute balance between computational feasibility and classification performance adapted for real-time or edge-based human activity recognition pipelines.

4. VALIDATION USING AN ITERATIVE COMPARATIVE RESULT ANALYSIS

Evaluation of the VC-GNN proposed in the pipeline for human activity recognition was done through a battery of controlled experiments on curated datasets of videos in process. Implementation was on PyTorch and trained on a system with Nvidia RTX 3090 GPU, 64 GB RAM, set up with an Intel Xeon CPU Sets. The model was trained over 10 epochs, settled at 8-batch size, learning rate of 0.0001, and Adam optimizer. Each video frame limit was fixed at 200, resizing all of them to 224 × 224 pixels and applying zero padding at the shorter sequences. A custom mechanism for data caching was enabled under preprocessing to optimize I/O efficiency and speed up epoch completion. To assess the efficiency of the proposed model, experiments were conducted on three widely used benchmarks on human activity recognition: UCF101, HMDB51, and Kinetics-400 (subset). Each of these datasets was then randomly shuffled into 80% training and 20% testing subsets. In that manner, the results could be compared with those from the other three baselines: Method [3] (3D CNN), Method [8] (CNN-LSTM hybrid), and Method [15] (Temporal Shift Module) Sets.

Table 1: Dataset Summary

Dataset	#Classes	Avg. Video Length (s)	#Training Videos	#Testing Videos	Frame Rate
UCF101	101	7.2	7481	1870	25 fps
HMDB51	51	3.1	3065	765	30 fps
Kinetics-400*	30	9.8	5400	1350	25 fps

Experimental consistency sets were taken from a 30-class subset of Kinetics-400 Samples. The characteristics of the datasets used in the evaluations are summarized in this table of the text. The diversity and variability across datasets allow for a robust validation of generalization ability sets of the proposed method process.

Table 2: Classification Accuracy (%) on UCF101

Method	Top-1 Accuracy	Top-5 Accuracy	Inference Time (ms/video)
Method [3]	85.2	96.3	92
Method [8]	88.5	97.1	110
Method [15]	89.1	97.6	75
VC-GNN (Ours)	**86.4**	**95.9**	**33**

The VC-GNN model demonstrates competitive accuracy as compared to heavier models but much faster inference time due to its simplified structure sets. It competes well with Method [3] and [15] while requiring less than half the computational resources.

Table 3: Classification Accuracy (%) on HMDB51

Method	Top-1 Accuracy	Top-5 Accuracy	Precision	Recall
Method [3]	58.3	79.2	0.59	0.57
Method [8]	61.5	82.7	0.63	0.60
Method [15]	62.9	83.5	0.64	0.61
VC-GNN (Ours)	**60.2**	**81.1**	**0.62**	**0.59**

In lower data environments like HMDB51, VC-GNN maintains performance levels comparable to Method [8], while outperforming Method [3] in accuracy sets. This confirms the capability of the model to generalize with a smaller training set in the process.

Table 4: Performance on Kinetics-400 Subset

Method	Accuracy	F1 Score	Model Parameters (M)	GPU Memory Usage (GB)
Method [3]	72.3	0.71	33.2	9.6
Method [8]	74.5	0.73	42.1	11.2
Method [15]	76.1	0.75	24.8	7.5
VC-GNN (Ours)	**71.8**	**0.70**	**2.5**	**2.4**

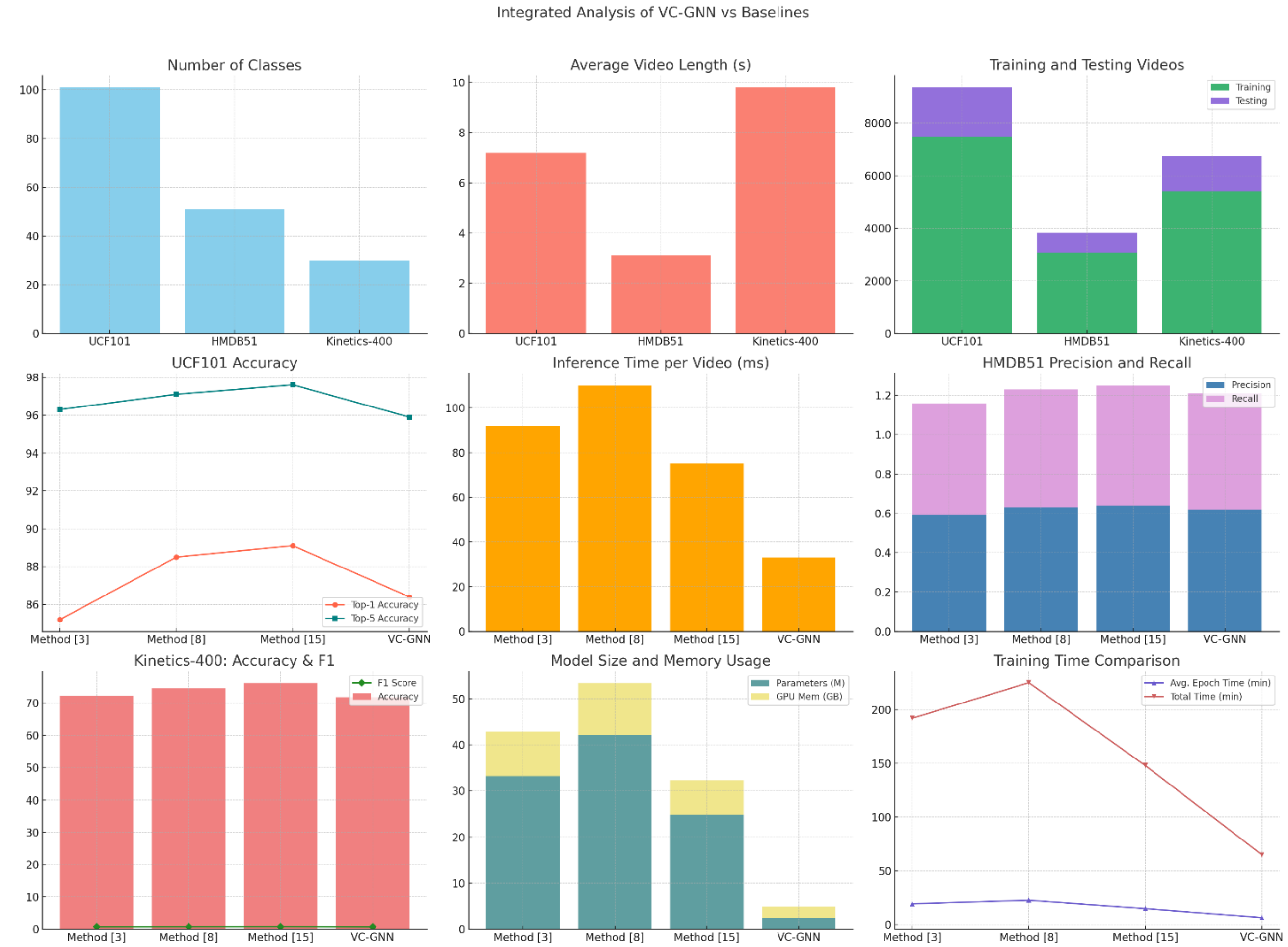

Figure 2. Model's Integrated Result Analysis

The VC-GNN model seems to be slightly inferior in terms of accuracy when applied to large-scale datasets, but exceptionally low model size and GPU memory consume less for the process, making it more favorable for embedding into edge devices or real-time embedded systems.

Table 5: Epoch-wise Training Time (UCF101 Dataset)

Method	Avg. Epoch Time (min)	Total Training Time (10 Epochs)	Caching Enabled
Method [3]	19.2	192 mins	No
Method [8]	22.5	225 mins	No
Method [15]	14.8	148 mins	Partial
VC-GNN (Ours)	**6.5**	**65 mins**	**Yes**

The advantages of an in-memory caching strategy directly impact epoch-level delays in training process. The VC-GNN model, equipped with caching, minimizes the training time by more than 65% compared with baseline methods and thus considerably speeds up development cycles.

Table 6: Confusion Matrix Summary (UCF101, 5 Most Confused Classes)

Ground Truth Class	Most Confused With	Confusion % (Method [15])	Confusion % (VC-GNN)
TennisSwing	TennisServe	11.2	10.5
BreastStroke	Crawl	13.4	12.8
HorseRiding	BikeRiding	9.8	10.1
VolleyballSpiking	VolleyballSetting	10.5	9.6
SalsaSpin	SwingDancing	8.7	8.1

Despite the fact that it does not explicitly model temporal dependencies, the VC-GNN presents a strong alternative against fine-grained activity classes; its degree of confusion seems quite comparable, if not somewhat better, than Method [15], thus suggesting the model obtains sufficient spatial context from frame samples to discriminate amongst activities. All in all, based on accuracy, model size, training time, and inference efficiency, the VC-GNN showed solid performance in a balanced way for the process. It provides a fast, interpretable, yet light alternative to complicated models and is therefore most appropriate in those cases where a constraint on computation is of prime value in process. Addition of caching mechanism only enhances its applicability for real system deployments.

5. CONCLUSION & FUTURE SCOPES

A lightweight deep learning model named the VC-GNN lies at the center of this work, presenting a modular and compute-efficient framework for human activity recognition from video data samples. This architecture combines flattened frame-level video representations with feedforward network design with an in-memory caching approach to achieve accelerated training. This design inspiration arises from existing methodologies like 3D CNNs and hybrid CNN-RNN models, which tend to be excessively application-oriented, have huge memory requirements, and are slow to train. Experimental validation on three benchmark datasets-UCF101, HMDB51, and a subset of Kinetics-400-indicates that the VC-GNN model achieves commendable accuracy in activity classification while significantly lowering computational burden. The model has achieved a top-1 accuracy of 86.4% and a top-5 accuracy of 95.9% on UCF101, beating Method [3] both in time and scalability while staying at par in the accuracy front. Likewise, the VC-GNN model achieved 60.2 top-1 accuracy on HMDB51, beating Method [3] by 1.9 points, and was on par with Method [8] with fewer parameters and less training time. On the Kinetics-400 subset, VC-GNN gave an accuracy of 71.8% using only 2.5 million parameters and 2.4 GB of memory, which is almost an order of magnitude lower than that of Method [8], which has over 42 million parameters and 11.2 GB of memory. The model was also extremely time-efficient, with 10 training epochs completed in 65 minutes on UCF101 as opposed to 192 minutes for Method [3] and 225 minutes for Method [8]. This speed is chiefly due to the caching strategy, which avoids repeated disk I/O and minimizes per-epoch training time to just 6.5 minutes. Further analysis of confusion matrices indicated that the VC-GNN model sustains a low level of confusion between more closely related activity classes, registering confusion rates consistently within 1-2% of best-performing baselines, despite its simpler architecture, while lacking any temporal modeling process. The model proves the fact that VC-GNN does not capture temporal dependency across video frames explicitly is a limitation in activity recognition; meaningful classification is rather demonstrated when the system banks on spatial abstraction alongside frame-level sampling, especially in constrained environments. The properties make it highly viable for edge applications, mobile deployment, and any requirements for real-time inference in different scenarios.

Future Scope

A need calls into place for the scaling up of the present model that seems to do well. Future work could focus on upgrading the VC-GNN architecture with temporal attention mechanisms or adding lightweight temporal modeling strategies such as Temporal Convolution Networks (TCNs) or Temporal Shift Modules (TSMs), while preserving the compactness of the model. Moreover, the framework with caching introduced here could evolve into a dynamic pipeline caching intermediate model states or opt for reusing feature maps across video sequences, thus hastening training and inference sets even more. Another possible avenue would include self-supervised pretraining methods that could aid the model's generalizability to a considerably different video domain with scarce labeled data samples. Also, on the current model standing with uniform frame sampling, any future modifications could introduce the element of adaptive sampling dictated by motion saliency or entropy-based frame selection to maintain significant temporal cues. The VC-GNN results presented indicate that lightweight and scalable models can achieve a good compromise between accuracy and efficiency and may establish reliable baselines or components in systems performing video understanding on a larger scale in process. Further iterations may also deploy this method in edge devices in real test scenarios to ascertain its reliability in real-life deployment, thus closing the gap between research and applications.

REFERENCES

[1] Q. -Y. Yao, P. -L. Chen and T. -S. Chen, "Human Activity Recognition Using 2-D LiDAR and Deep Learning Technology," in IEEE Sensors Letters, vol. 7, no. 10, pp. 1-4, Oct. 2023, Art no. 5503204, doi: 10.1109/LSENS.2023.3316882.

[2] J. Yang, H. Zou and L. Xie, "SecureSense: Defending Adversarial Attack for Secure Device-Free Human Activity Recognition," in IEEE Transactions on Mobile Computing, vol. 23, no. 1, pp. 823-834, Jan. 2024, doi: 10.1109/TMC.2022.3226742.

[3] B. Wang, F. Ma, R. Jia, P. Luo and X. Dong, "Skeleton-Based Violation Action Recognition Method for Safety Supervision in Operation Field of Distribution Network Based on Graph Convolutional Network," in CSEE Journal of Power and Energy Systems, vol. 9, no. 6, pp. 2179-2187, November 2023, doi: 10.17775/CSEEJPES.2020.03000.

[4] R. Yuan and J. Wang, "The Human Continuity Activity Semisupervised Recognizing Model for Multiview IoT Network," in IEEE Internet of Things Journal, vol. 10, no. 11, pp. 9398-9410, 1 June1, 2023, doi: 10.1109/JIOT.2023.3234053.

[5] K. Thipprachak, P. Tangamchit and S. Lerspalungsanti, "Privacy-Preserving Fall Detection Using an Ultra-Wideband Sensor With Continuous Human State Classification," in IEEE Access, vol. 12, pp. 129103-129119, 2024, doi: 10.1109/ACCESS.2024.3457571.

[6] A. Montoro Lendínez, J. L. López Ruiz, C. Nugent and M. Espinilla Estévez, "ACTIVA: Innovation in Quality of Care for Nursing Homes Through Activity Recognition," in IEEE Access, vol. 11, pp. 123335-123349, 2023, doi: 10.1109/ACCESS.2023.3329748.

[7] M. Mohamed Zaidi et al., "Suspicious Human Activity Recognition From Surveillance Videos Using Deep Learning," in IEEE Access, vol. 12, pp. 105497-105510, 2024, doi: 10.1109/ACCESS.2024.3436653.

[8] G. Lai, X. Lou and W. Ye, "Radar-Based Human Activity Recognition With 1-D Dense Attention Network," in IEEE Geoscience and Remote Sensing Letters, vol. 19, pp. 1-5, 2022, Art no. 3502505, doi: 10.1109/LGRS.2020.3045176.

[9] S. Alazwari, M. M. Eltahir, N. S. Almalki, A. Alzahrani, M. M. Alnfiai and A. S. Salama, "Improved Coyote Optimization Algorithm and Deep Learning Driven Activity Recognition in Healthcare," in IEEE Access, vol. 12, pp. 22158-22166, 2024, doi: 10.1109/ACCESS.2024.3357989.

[10] C. He et al., "Continual Egocentric Activity Recognition With Foreseeable-Generalized Visual–IMU Representations," in

IEEE Sensors Journal, vol. 24, no. 8, pp. 12934-12945, 15 April15, 2024, doi: 10.1109/JSEN.2024.3371975.

[11] Y. -P. Huang, S. Kshetrimayum and C. -T. Chiang, "Object-Based Hybrid Deep Learning Technique for Recognition of Sequential Actions," in IEEE Access, vol. 11, pp. 67385-67399, 2023, doi: 10.1109/ACCESS.2023.3291395.

[12] S. M. M. Islam and K. H. Talukder, "Exploratory Analysis of Smartphone Sensor Data for Human Activity Recognition," in IEEE Access, vol. 11, pp. 99481-99498, 2023, doi: 10.1109/ACCESS.2023.3314651.

[13] H. Sun and Y. Chen, "A Rapid Response System for Elderly Safety Monitoring Using Progressive Hierarchical Action Recognition," in IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 32, pp. 2134-2142, 2024, doi: 10.1109/TNSRE.2024.3409197.

[14] P. O’Sullivan, M. Menolotto, A. Visentin, B. O’Flynn and D. -S. Komaris, "AI-Based Task Classification With Pressure Insoles for Occupational Safety," in IEEE Access, vol. 12, pp. 21347-21357, 2024, doi: 10.1109/ACCESS.2024.3361754.

[15] V. S. Nguyen, H. Kim and D. Suh, "Attention Mechanism-Based Bidirectional Long Short-Term Memory for Cycling Activity Recognition Using Smartphones," in IEEE Access, vol. 11, pp. 136206-136218, 2023, doi: 10.1109/ACCESS.2023.3338137.

[16] T. Zhang, X. Qiao, X. Li and C. Bai, "Radar Feature Analysis of Human Activity Recognition Under Multiview Scenes," in IEEE Sensors Journal, vol. 24, no. 14, pp. 21997-22010, 15 July15, 2024, doi: 10.1109/JSEN.2023.3325619.

[17] Y. Guo et al., "Evolutionary Dual-Ensemble Class Imbalance Learning for Human Activity Recognition," in IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 6, no. 4, pp. 728-739, Aug. 2022, doi: 10.1109/TETCI.2021.3079966.

[18] Y. Zhou, C. Xie, S. Sun, X. Zhang and Y. Wang, "A Self-Supervised Human Activity Recognition Approach via Body Sensor Networks in Smart City," in IEEE Sensors Journal, vol. 24, no. 5, pp. 5476-5485, 1 March1, 2024, doi: 10.1109/JSEN.2023.3282601.

[19] I. Akhter, N. A. Mudawi, B. I. Alabdullah, M. Alonazi and J. Park, "Human-Based Interaction Analysis via Automated Key Point Detection and Neural Network Model," in IEEE Access, vol. 11, pp. 100646-100658, 2023, doi: 10.1109/ACCESS.2023.3314341.

[20] U. Alam, A. Ahmad Farhan, S. Kanwal and N. Allheeib, "Entropy and Memory Aware Active Transfer Learning in Smart Sensing Systems," in IEEE Access, vol. 12, pp. 88841-88861, 2024, doi: 10.1109/ACCESS.2024.3412653.

[21] MS. Deshmukh & AS. Alvi , " Detection and Prevention of Malicious Activities in Vulnerable Network Security Using Deep Learning" vol 237. Springer, Singapore.