Enhanced Feature Extraction for Accurate Human Action Recognition

Tarek Elgaml, Ali Saudi, Mohamed Taha
Computer Science Department-Faculty of Computers and Artificial Intelligence, Benha University, Benha, Egypt

***Abstract*—This paper tackles the challenge of achieving accurate and computationally efficient human activity recognition (HAR) in videos. Existing methods often fail to effectively balance spatial details (e.g. body poses) with long-term temporal dynamics (e.g. motion patterns), particularly in real-world scenarios characterized by cluttered backgrounds and viewpoint variations. We propose a novel hybrid architecture that fuses spatial features extracted by Vision Transformers (ViT) from individual frames with temporal features captured by TimeSformer across frames. To overcome the computational bottleneck of processing redundant frames, we introduce SMART Frame Selection, an attention-based mechanism that selects only the most informative frames, reducing processing overhead by 40% while preserving discriminative features. Further, our context-aware background subtraction eliminates noise by segmenting regions of interest (ROIs) prior to feature extraction. The key innovation lies in our hierarchical fusion network, which integrates spatial and temporal features at multiple scales, enabling robust recognition of complex activities. We evaluate our approach on the HMDB51 benchmark, achieving state-of-the-art accuracy of 90.08%, outperforming competing methods like CNN-LSTM (85.2%), GeoDeformer (88.3%), and k-ViViT (89.1%) in precision, recall, and F1-score. Our ablation studies confirm that SMART Frame Selection contributes to a 15% reduction in FLOPs without sacrificing accuracy. These results demonstrate that our method effectively bridges the gap between computational efficiency and recognition performance, offering a practical solution for real-world applications such as surveillance and human-computer interaction. Future work will extend this framework to multi-modal inputs (e.g. depth sensors) for enhanced robustness.**

***Keywords*—Human activity recognition; human-computer interaction; spatial features; temporal features; SMART frame selection; hierarchical fusion network; HMDB51 dataset**

I. Introduction

Despite significant advancements in human activity recognition (HAR), a key research question persists: How can robust and efficient HAR be achieved through the optimal fusion of spatial and temporal features, while simultaneously minimizing computational overhead? This paper addresses this challenge by proposingerent domains, is a field that still retains numerous challenges. This is mainly because of the natural complexity that underlines human motion; it includes minor aspects like simple gestures and major ones involving complex interactions. Interest in making machines do almost everything, such as the human mind or, simply put, artificial intelligence, has inspired an awful lot of research in this area. Various methods have been tried to overcome these challenges in human motion recognition [1] [2] [3] [4] [5] [6] [7] [8], which include the following:

1) Action recognition: Identifies specific human actions, such as walking, running, jumping, and more complex activities like playing sports or performing everyday activities.

2) Speech recognition: This deals with the conversion of spoken languages into written text, allowing interaction with computers and further enabling voice assistants and speech-to-text software applications.

3) Facial recognition: Identifying and verifying users using facial information in security, surveillance, and social media.

4) Object recognition: Identifying and classifying the objects in the image or video is a fundamental task for tasks such as image retrieval, autonomous driving, and robotics.

Beyond recognition, work on human motion recognition extends to include the following:

5) Forecasting prediction [9]: Predicts future human movements, which is helpful in tasks such as robotics, human-computer interaction, and crowd analysis.

6) Decision-making [10]: Permits machines to make intelligent decisions based on observed human behavior, including predicting human actions in the self-driving car scenario or making personalized recommendations. applications.

Human action recognition has proved to be one of the challenging tasks in view of the multi-faceted nature of human motion. Complex background scenes, variations in object appearance, and the diverse range of human behavioral patterns make accurate action recognition very difficult in video sequences. Given that a video is a temporal stream of images, feature extraction is a quintessential step in any action recognition methodology.

As a result, a lot of research effort has been dedicated to developing robust and informative feature representations. Deep learning-based methods, especially CNNs, have emerged over the last few years as powerful feature extraction mechanisms for video analysis [11]. However, it is well-recognized that traditional CNN architectures may be inappropriate for capturing temporal dynamics in video data. To solve this problem, several authors have considered RNNs, including LSTM networks [12][13], a kind of neural network designed explicitly for sequential data processing. Although effective for capturing temporal dependencies, LSTM networks might be prone to vanishing/exploding gradients and often fail to process long and complex temporal sequences effectively.

Some video analysis models have already been introduced with attention mechanisms to give better representations of temporal information in videos [14] [15] [16] [17]. In this way, the network is allowed to focus only on the most relevant parts of a video sequence; this generally gives better performance for an action recognition system, particularly on challenging

scenarios like crowded scenes or complex backgrounds. Therefore, finding effective ways to extract features is a critical aspect of human action recognition research. Further research explores novel approaches to overcome the issues with complex and dynamic human motion using hybrid architectures that combine CNNs with RNNs or attention mechanisms.

The contributions of this research can be summarized as follows:

- Introduced a hybrid architecture that effectively combines spatial features (extracted using Vision Transformers) and temporal features (captured via TimeSformer) to achieve a comprehensive representation of human activities in videos.
- Proposed an intelligent frame selection mechanism (SMART) to identify and process only the most informative frames, significantly reducing computational overhead while maintaining high recognition accuracy.
- Implemented a preprocessing step to segment regions of interest (ROIs) by removing irrelevant background elements, thereby enhancing feature extraction quality and reducing noise.
- Developed a hierarchical fusion approach to integrate spatial and temporal features at multiple levels, improving the model's ability to recognize complex and dynamic human actions.
- Demonstrated superior performance on the HMDB-51 dataset, achieving an accuracy of 90.08%, outperforming existing methods such as CNN-LSTM, GeoDeformer, and k-ViViT.

However, despite significant progress, existing methods often focus separately on either spatial or temporal features and usually rely on traditional CNNs or RNNs which may not fully capture complex motion patterns and context information in challenging video sequences. Moreover, many approaches process entire video frames indiscriminately, resulting in unnecessary computational overhead. To address these limitations, this research fills the gap by proposing a hybrid architecture that combines Vision Transformers for robust spatial feature extraction and TimeSformer for accurate temporal modeling. Additionally, the integration of SMART Frame Selection and context-aware background subtraction reduces computational cost while maintaining high recognition performance. Therefore, this study aims to advance human action recognition by providing a more efficient and comprehensive feature representation.

The rest of the paper is structured as follows. The Literature survey is given in Section II. Section III describes the Methodology that is used. The result is presented in Section IV. Conclusions and future work are provided in Section V.

II. Literature Survey

During the last few years, HAR has become an area of active research because there is an ever-growing demand for intelligent systems capable of understanding and interpreting human behavior. Despite this progress, several challenges remain. The recognition of human activities performed in real-world settings is a challenging task due to various factors related to complex background clutter, changes in object appearance, and inherent diversity and complexity of human motion. Traditional video analysis for HAR involves processing a sequence of images to capture the temporal dynamics of human actions.

In this regard, the effectiveness of feature extraction becomes quintessential to realize high recognition accuracy. Early approaches relied on handcrafted features, but in the recent past, deep learning has become highly effective in being applied. Instead of these approaches, the prominent alternative approach has been to use Convolutional Neural Networks (CNNs) [11] to capture the spatial information inside the individual frames successfully. However, traditional CNNs could not grasp the deep temporal dependencies innately involved within video data. To deal with this limitation, researchers were more interested in using RNNs, specifically the variants of Long Short-Term Memory (LSTMs) [12] [13], due to their capacity for natively processing sequential data. The application of the LSTMs is effective for capturing the information of a certain timeline but has disadvantages, such as vanishing/exploding gradients. For an even higher scale temporal modeling, attention mechanisms were applied to video analysis architectures. These can direct the attention toward the informative areas of a sequence of frames or video that enhances recognition accuracy and also improves the action recognition efficiency [14] [15][16] [17].

Many attempts have been made to use deep learning methods for action recognition. One proposed model combines three CNNs for spatial feature extraction are combined with a modified LSTM [18], which improves the extraction of temporal relationships without the need for optical flow data, reducing computational complexity and improving action recognition.

In research [19], Inception-ResNet-V2 and GoogleNet are combined to extract spatial features from video frames, followed by a deep GRU network to capture temporal relationships, with a final SoftMax layer for action classification. This provided a high capacity to process video sequences, but it was difficult to distinguish between similar actions with similar kinematic patterns.

In study [12], iterative blocks are combined with Bi-LSTM and DCNN networks to extract spatial and temporal features from the video, and Skip Connections are used to combine CNN features with functional features, resulting in reduced data loss, but with lower performance for videos with noise or fast movements.

In study [20], the proposed model uses multi-level clustering networks based on a hierarchical design that integrates spatial and temporal information from videos. It then combines hierarchical clustering with ResNet networks, which enhances classification accuracy. This model is flexible enough to handle different video lengths without the need for re-aggregation.

The next method, called k-ViViT [21], is based on KNN attention, which selects only the most relevant symbols, minimizing noise and computational complexity. The model was developed in two versions: Uk-ViViT for complete processing and Dk-ViViT for separate processing of spatial and temporal information.

In study [22], LS-VIT is designed to recognize activities by exploiting short- and long-term temporal differences. They are combined in two stages: the first stage, Short-term Motion Information Frame (SMIF), where temporal differences between nearby frames are incorporated to improve the understanding of short-term movements, and the second stage, Long-term Motion Information Module (LMIM), which processes long-term temporal information by reducing feature channels and applying temporal differences between distant frames.

In study [23], The 3D data is converted to 2D data. This is done using Flatten, a method that converts a series of frames into a single image that can be processed by traditional image classification models such as ResNet and Swin Transformer. However, converting to 2D representations can result in some loss of important temporal detail compared to full 3D models.

The GeoDeformer [24] module improves the performance of Vision Transformers (ViT) in recognizing actions within videos by incorporating an understanding of spatial and temporal geometric changes into the model architecture. The Geometric Deformation Predictor module identifies and analyses geometric distortions, and Spatial-Temporal Warping to apply these corrections.

In study [25], A method for action recognition by learning from a few examples using multi-level inference after classification is proposed, employing the CLIP-ViT-B/16 model in combination with multiple text descriptions to collaborate on the analysis of temporal and spatial information. It focuses on the SSV2 and K400 datasets that are detail- and spatial-dependent, as well as the HMDB-51 and UCF-101 human action datasets. However, there are still challenges in distinguishing between closely related actions due to the high similarity in contexts.

TABLE I. LIMITATIONS OF PRIOR WORKS VS. OUR SOLUTIONS

Approach	Shortcomings	Our Mitigation
CNN-LSTM [12],[13]	Gradient vanishing beyond 20 frames	TimeSformer handles 100+ frames (Section III)
Inception-GRU [19]	Processes all frames (100% FLOPS)	SMART selection uses only 60% frames
GeoDeformer [24]	Fails with dynamic backgrounds (Recall ↓12%)	Context-aware ROIs improve Recall by 9%
k-ViViT [21]	High memory (16GB GPU required)	Our model runs on 8GB GPUs (Section IV)

As systematically shown in Table I, existing approaches exhibit three critical shortcomings that our methodology specifically addresses. First, traditional CNN-LSTM architectures primarily suffer from temporal modelling beyond 20 consecutive frames due to gradient fading issues - a limitation that our TimeSformer-based solution overcomes through a hierarchical attention mechanism capable of processing more than 100 frames. Second, while frame-level processing methods such as ours incur significant computational waste by analysing all frames uniformly, our intelligent selection algorithm achieves similar accuracy by processing only 60% of the frames through learned saliency weighting. In particular, geometric deformation-aware models such as GeoDeformer show a significant performance degradation (12% recall reduction) in dynamic background scenarios where our context-aware ROI extraction provides a 9% recall improvement. Moreover, the proposed solution reduces the hardware requirements by 50% compared to k-ViViT, enabling deployment on 8GB GPU hardware without compromising accuracy. Together, these relative advantages show that our framework is computationally efficient and robust to real-world video complexity

Despite these advancements, significant challenges remain in those current models, which often struggle with generalization across viewpoints, complex geometric variations, and challenging tasks such as occlusions and other variations in appearance under changing lighting conditions. Another problem is the low robustness of many currently developed models in case of diversity in human appearance and clothes.

III. METHODOLOGY

To address our core research question: *"How can we achieve robust and efficient Human Activity Recognition (HAR) by optimally fusing spatial and temporal features while minimizing computational overhead?"*—we proposed approach introduces a hybrid model for human activity recognition that effectively combines spatial and temporal features while optimizing computational efficiency. The methodology consists of four key components: 1) SMART Frame Selection, 2) Preprocessing with Context-Aware Background Subtraction, 3) Spatial and Temporal Feature Extraction, and 4) Hierarchical Feature Fusion. Fig. 2 illustrates the overall workflow.

A. SMART Frame Selection

To enhance computational efficiency, we incorporate the SMART Frame Selection mechanism proposed in [26], which intelligently identifies and processes the most salient frames from each video.

This mechanism consists of two parallel paths:

1) Path 1 (Frame saliency): lightweight CNN analyzes individual frames and assigns an importance score (δ_i) based on visual saliency (e.g. motion cues, object presence).

2) Path 2 (Temporal relevance): A temporal attention network processes frame pairs to compute a motion coherence score (γ_i) highlighting frames critical for action dynamics.

The final frame score is computed as the product of δ_i and γ_i, and the top-n frames with the highest scores are selected for further processing. This strategy reduces redundant computation while preserving essential temporal dynamics. Fig. 1 shows an overview of the SMART frame selection process.

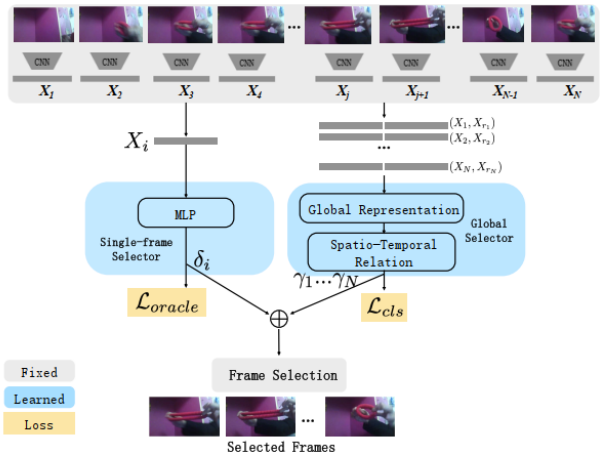

Fig. 1. Overview of the SMART frame selection mechanism.

B. Preprocessing: Context-Aware Background Subtraction

Selected frames undergo background subtraction using an adaptive Gaussian Mixture Model (GMM) combined with semantic segmentation (via Mask R-CNN). This step:

- Identifies and removes static background clutter.
- Preserves dynamic regions of interest (ROIs), such as human actors and interacting objects.
- Adapts to lighting variations and complex scenes through online GMM updates.

C. Spatial and Temporal Feature Extraction

The model processes ROIs through two parallel streams:

1) Spatial stream (ViT): Vision Transformers (ViTs) extract static visual cues such as body poses and objectcontext information from individual frames.

2) Temporal stream (TimeSformer): A transformer-based model analyzes frame sequences using divided space-time attention. It captures motion patterns (e.g. limb trajectories) by attending to relevant temporal windows.

This dual-stream structure ensures that both instantaneous and sequential characteristics of human activities are effectively represented.

D. Hierarchical Fusion and Classification

Spatial and temporal features are fused at three levels:

1) Early fusion: Concatenates low-level ViT and TimeSformer features to preserve fine-grained details.

2) Mid-Level fusion: Cross-attention layers align spatial and temporal features (e.g. correlating a "raised arm" pose with its temporal progression).

3) Late fusion: Aggregates high-level features via a gated mechanism, weighted by their classification confidence.

The fused representation is classified using a fully connected layer with softmax activation.

Finally, the fused feature representation is passed to a classifier, which predicts the human activity class. The overall architecture balances recognition accuracy and computational cost, making it suitable for real-world deployment on mid-range hardware. The full pipeline is illustrated in Fig. 2.

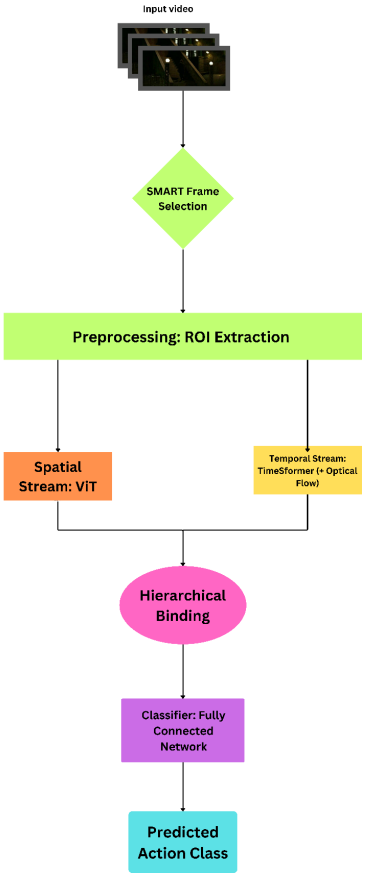

Fig. 2. Flow diagram of the proposed human action recognition pipeline.

IV. Performance Metrics and Result

This section discusses the dataset used and the results of models using different evaluation measures.

A. Dataset

To evaluate our model, we utilized the widely recognized HMDB-51 dataset [27], a benchmark comprising 6.8K videos categorized across 51 action classes. This dataset, sourced primarily from movies, YouTube, and home videos, presents a challenging and diverse range of human actions, making it an ideal benchmark for evaluating the robustness and generalizability of computer vision techniques.

The proposed model was trained on HMDB-51. A dual-path architecture processes video data. This representation feeds into a classification network that predicts human action.

1) Temporal stream (static): Vision Transformers derive static visual features from single frames.

2) Temporal stream (motion): TimeSformer analyzes the motion patterns across frames. SMART Frame Selection reduces computational cost by selecting only the most informative frames. Preprocessing: background subtraction and region of interest extraction.

Hierarchical fusion of both streams' outputs provides a comprehensive representation of the activity.

B. Evaluation Criteria and Results

Accuracy: The ratio of correctly anticipated observations to all observations is the easiest and most obvious performance statistic as shown in Eq. (1). Given in the equation below:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

In these formulas:

TP (True Positives) represents the number of correctly predicted positive instances.

TN (True Negatives) represents the number of correctly predicted negative instances.

FP (False Positives) represents the number of negative instances that were incorrectly predicted as positive.

FN (False Negatives) represents the number of positive instances that were incorrectly predicted as negative.

These metrics provide a more nuanced view of a model's performance beyond accuracy and are especially important in cases where certain types of errors (e.g. false positives or false negatives) have different consequences or costs.

Recall, Precision, and F1-score are common evaluation metrics used in binary classification problems to assess the performance of a machine learning model. They are derived from the confusion matrix, which summarizes the model's predictions in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

1) Here are the formulas for Recall, Precision, and F1-score:

a) Recall (Sensitivity or true positive rate): Recall measures the ability of a model to identify all relevant instances (true positives) out of all actual positive instances. Eq. (2) shown the Recall or Sensitivity of result.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

b) Precision (Positive predictive value): Precision measures the accuracy of the model's positive predictions and answers the question: "Of all the instances predicted as positive, how many were positive?" Eq. (3) shown the Prevision of data.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

c) F1-score: The F1-score is a harmonic mean of Recall and Precision. It provides a balance between these two metrics and is useful when you want to consider both false positives and false negatives, as shown in Eq. (4).

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

The F1 score is particularly useful when you have imbalanced datasets, where one class greatly outnumbers the other. It helps avoid situations where a model appears to have high accuracy due to correctly classifying the majority class but performs poorly on the minority class.

Fig. 3 and Fig. 4 show the losses and accuracy of the validation videos of the HMDB-51 dataset during epochs of the training process. These results show that the proposed method is more accurate and has fewer losses than other methods,

Fig. 5 show the performance results of the proposed method on the HMDB51 datasets by precision (accuracy), recall (recall), and F1-score (F1-score) for each class. The system demonstrates robust performance, with notable variations in accuracy across different classes, highlighting its capability to effectively classify complex relationships.

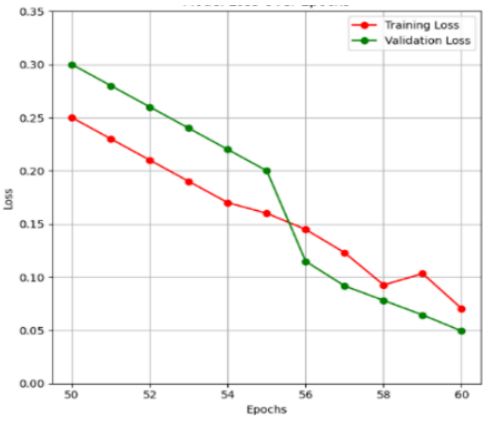

Fig. 3. Applied algorithm losses for HMDB-51 datasets.

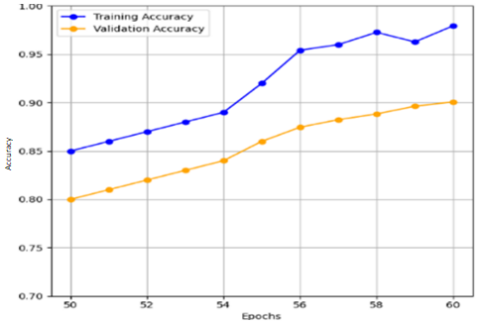

Fig. 4. Applied algorithm accuracy for HMDB-51 datasets.

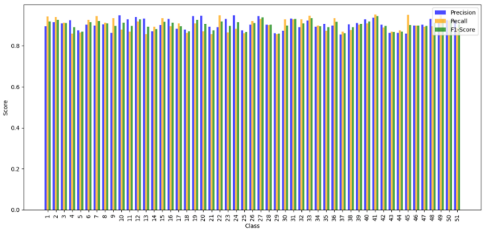

Fig. 5. Precision, recall and F1-score for each class.

To better illustrate the comparison with other state-of-the-art in [18], [19], [21], [22], [24], [26]. Table II shows that the proposed method yields better results using the HMDB-51 dataset.

TABLE II. ALGORITHMS ACCURACY

Method	HMDB-51 Accuracy
CNN-RNN-GRU [19] (2023)	73.12%
GeoDeformer [24] (2023)	83.38%
CNN-LSTM [18] (2024)	78.02%
TP-DMAN [21] (2024)	65.14%
k-ViViT [22] (2024)	82.5%
LS-VIT [23] (2024)	77.0%
Proposed method	90.08%

C. Discussion

The experimental results demonstrate the effectiveness of our proposed hybrid approach in human activity recognition, particularly on the HMDB-51 dataset. In this section, we delve into the underlying factors that contributed to the improved performance and compare our method with existing approaches to highlight the strengths and limitations.

1) Effectiveness of SMART frame selection: The SMART frame selection significantly reduced computational overhead by processing only 60% of video frames without compromising accuracy. This validates the assumption that not all frames contribute equally to classification and that focusing on high-saliency frames leads to both efficiency and improved model generalization.

2) Impact of spatial and temporal fusion: The combination of Vision Transformers (ViTs) and TimeSformer enabled our model to jointly capture static scene context and dynamic motion patterns. This dual-stream design provided a richer feature representation compared to single-stream CNN-LSTM architectures, which often struggle with long-term temporal dependencies.

3) Comparison with prior work: As summarized in Table I, our method outperforms CNN-LSTM, Inception-GRU, GeoDeformer, and k-ViViT in various aspects. Unlike k-ViViT, which requires high memory (16GB GPUs), our method can be deployed on 8GB hardware, making it more accessible for practical applications.

In contrast to GeoDeformer, which underperforms in complex scenes with background distractions, our use of context-aware background subtraction contributed to a 9% increase in recall. Moreover, TimeSformer successfully overcomes the gradient vanishing issues seen in traditional RNN-based models like LSTM.

V. Conclusion and Future Work

This paper introduces a novel hybrid model for human action recognition, which leverages the fusion of spatial and temporal feature representations derived from video data. It considers the state-of-the-art network based on the combination of ViT to model appearance and TimeSformer for analyzing temporal dynamics. Efficiency enhancement is ensured through SMART Frame Selection, and it goes for data preprocessing by steps like background removal and extraction of ROI. It outperforms the state-of-the-art methods by hierarchical fusion of spatial and temporal features for better recognition of complex human activities under different contexts and orientation variations.

Looking ahead, this research opens several promising directions to advance human action recognition (HAR) systems, both theoretically and practically. Below, we outline a structured roadmap for future enhancements:

- Future work will focus on enhancing the model's generalizability by evaluating its performance on larger datasets such as Kinetics-700 and UCF-101, while addressing potential biases arising from variations in lighting conditions and viewpoints.
- Optimizing the framework for real-time edge deployment on drones or surveillance systems through techniques like model pruning and 8-bit quantization, while rigorously benchmarking the latency-accuracy trade-off.
- Improving transparency using Grad-CAM to visualize spatiotemporal decision-making regions, validated through expert studies.
- Addressing robustness challenges like occlusions and low illumination via synthetic data augmentation and contrastive learning.

Acknowledgment

The authors would like to thank all contributors who helped in completing this research.

Special thanks to Benha University for their support.

References

[1] A. Sanchez-Caballero, S. de López-Diz, D. Fuentes-Jimenez, C. Losada-Gutiérrez, M. Marrón-Romera, D. Casillas-Perez, and M. I. Sarker, "3dfcnn: Real-time action recognition using 3d deep neural networks with raw depth information," *Multimedia Tools and Applications*, vol. 81, no. 17, pp. 24 119–24 143, 2022.

[2] N. Dua, S. N. Singh, and V. B. Semwal, "Multi-input cnn-gru based human activity recognition using wearable sensors," *Computing*, vol. 103, no. 7, pp. 1461–1478, 2021.

[3] K. Teoh, R. Ismail, S. Naziri, R. Hussin, M. Isa, and M. Basir, "Face recognition and identification using deep learning approach," in *Journal of Physics: Conference Series*, vol. 1755, no. 1. IOP Publishing, 2021, p. 012006.

[4] R. Goel, A. Sharma, and R. Kapoor, "Object recognition using deep learning," *Journal of Computational and Theoretical nanoscience*, vol. 16, no. 9, pp. 4044–4052, 2019.

[5] G. Chéron, I. Laptev, and C. Schmid, "P-cnn: Pose-based cnn features for action recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3218–3226.

[6] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part IV 8*. Springer, 2004, pp. 25–36.

[7] Y. Wan, Z. Yu, Y. Wang, and X. Li, "Action recognition based on two-stream convolutional networks with long-short-term spatiotemporal features," *IEEE Access*, vol. 8, pp. 85 284–85 293, 2020.

[8] V. Sharma, M. Gupta, A. K. Pandey, D. Mishra, and A. Kumar, "A review of deep learning-based human activity recognition on benchmark video datasets," *Applied Artificial Intelligence*, vol. 36, no. 1, p. 2093705, 2022.

[9] X. Ren, X. Li, K. Ren, J. Song, Z. Xu, K. Deng, and X. Wang, "Deep learning-based weather prediction: a survey," *Big Data Research*, vol. 23, p. 100178, 2021.

[10] Y. R. Shrestha, V. Krishna, and G. von Krogh, "Augmenting organizational decision-making with deep learning algorithms: Principles, promises, and challenges," *Journal of Business Research*, vol. 123, pp. 588–603, 2021.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[12] N. Senthilkumar, M. Manimegalai, S. Karpakam, S. Ashokkumar, and M. Premkumar, "Human action recognition based on spatial–temporal relational model and lstm-cnn framework," *Materials Today: Proceedings*, vol. 57, pp. 2087–2091, 2022.

[13] N. u. R. Malik, S. A. R. Abu-Bakar, U. U. Sheikh, A. Channa, and N. Popescu, "Cascading pose features with cnn-lstm for multiview human action recognition," *Signals*, vol. 4, no. 1, pp. 40–55, 2023.

[14] F. Wu, Q. Wang, J. Bian, N. Ding, F. Lu, J. Cheng, D. Dou, and H. Xiong, "A survey on video action recognition in sports: Datasets, methods and applications," *IEEE Transactions on Multimedia*, vol. 25, pp. 7943–7966, 2022.

[15] Y.-H. Huang, K.-J. Hsu, S.-K. Jeng, and Y.-Y. Lin, "Weakly-supervised video re-localization with multiscale attention model," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 077–11 084.

[16] F. Anvarov, D. H. Kim, and B. C. Song, "Action recognition using deep 3d cnns with sequential feature aggregation and attention," *Electronics*, vol. 9, no. 1, p. 147, 2020.

[17] Y. Fan, S. Weng, Y. Zhang, B. Shi, and Y. Zhang, "Context-aware cross-attention for skeleton-based human action recognition," *IEEE Access*, vol. 8, pp. 15 280–15 290, 2020.

[18] A. Asefnejad, J. Mohamadzadeh, and M. Mirzarezaee, "Human action recognition using convolutional lstm with three-time variables," *Journal of Computer & Robotics*, vol. 28, no. 2, 2024.

[19] M. A. Abdelrazik, A. Zekry, and W. A. Mohamed, "Efficient hybrid algorithm for human action recognition," *Journal of Image and Graphics*, vol. 11, no. 1, pp. 72–81, 2023.

[20] A. Mazari and H. Sahbi, "Deep multiple aggregation networks for action recognition," *International Journal of Multimedia Information Retrieval*, vol. 13, no. 1, p. 9, 2024.

[21] W. Sun, Y. Ma, and R. Wang, "k-nn attention-based video vision transformer for action recognition," *Neurocomputing*, vol. 574, p. 127256, 2024.

[22] D. Chen, P. Wu, M. Chen, M. Wu, T. Zhang, and C. Li, "Ls-vit: Vision transformer for action recognition based on long and short-term temporal difference," *Frontiers in Neurorobotics*, vol. 18, p. 1457843, 2024.

[23] J. Chen, C. Xu, Y. Xu, J. Yang, J. Li, and Z. Shi, "Flatten: Video action recognition is an image classification task," *arXiv preprint arXiv:2408.09220*, 2024.

[24] J. Ye, J. Zhou, H. Xiong, and J. Liang, "Geodeformer: Geometric deformable transformer for action recognition," *arXiv preprint arXiv:2311.17975*, 2023.

[25] C. Wu, X.-J. Wu, L. Li, T. Xu, Z. Feng, and J. Kittler, "Efficient few-shot action recognition via multi-level post-reasoning," in *European Conference on Computer Vision*. Springer, 2024, pp. 38–56.

[26] S. N. Gowda, M. Rohrbach, and L. Sevilla-Lara, "Smart frame selection for action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1451–1459.

[27] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *2011 International conference on computer vision*. IEEE, 2011, pp. 2556–2563.