Kalpa Publications in Computing

Volume 19, 2024, Pages 82–93

Proceedings of 6th International Conference
on Smart Systems and Inventive Technology

The Enhanced Human Activity Recognition: A Context Based Two-Stream Framework for Learning Temporal and Spatial Features

Milind V. Kamble[1] and Rajankumar Bichkar[2]

[1] Research Scholar, Dept. of E&TC,
G. H. Raisoni College of Engineering
and Management, Pune, India
milind.kamble@vit.edu

[2] Vidya Pratishthan's Kamalnayan Bajaj
Institute of Engineering and Technology
Baramati, Maharashtra, India
bichkar@yahoo.com

Abstract

Human Activity Recognition (HAR) from video signals is increasingly crucial for surveillance, healthcare, robotics, and augmented reality applications. Accurately identifying human actions is vital in our data-driven world, posing a significant technological challenge. This study introduces a comprehensive methodology for HAR, starting with the preprocessing of video frames using context-awareness. The context-aware frames are then fed into a two-stream framework, extracting spatial and temporal features in a complementary manner. The spatial stream analyzes visual features from individual video frames, while the temporal stream focuses on dynamic aspects, capturing intricate motion patterns. This separation allows for a detailed analysis of video data, aligning with human perception of activities. The subsequent stage involves a late binding mechanism, enabling optimal interaction between spatial and temporal streams. Integration in a dense layer allows the model to harness interactions between these information streams, significantly improving recognition accuracy. Rigorous experimental validation confirms the efficacy and reliability of the proposed approach in diverse scenarios using real-world datasets HMDB51 and UCF50. The results demonstrate high accuracy, precision, recall, and F-measure for the combined spatial and temporal model compared to individual streams. This research contributes to advancing HAR technology, improving how computers interpret and recognize human activities in videos for practical and beneficial applications.

1 Introduction

Human activity recognition (HAR) from video signals is a field of growing significance with an array of practical applications. Understanding and accurately identifying human actions in video data are crucial for domains ranging from surveillance and healthcare to robotics and

augmented reality. As our world becomes increasingly data-driven, the ability to automatically interpret human activities has become a pivotal technological challenge [3]. In recent years, significant advancements have been made in HAR by applying deep learning techniques, improving the HAR system's accuracy and robustness. These improvements are primarily attributed to extracting spatial and temporal information from video data. Spatial information captures static visual cues, such as body posture, object interactions, and scene context, while temporal information characterizes the dynamic patterns and motion evolution over time. Combining these two types of information has led to a substantial leap in recognition performance [10]. Our work contributes to the evolving field of HAR by introducing a novel two-stream framework. This approach extracts spatial and temporal features complementary, enhancing detailed video data analysis. The spatial stream focuses on visual features from individual frames, while the temporal stream captures dynamic aspects, aligning with natural human perception. Our methodology incorporates a late binding mechanism, optimizing the interaction between spatial and temporal streams in a dense layer, significantly improving recognition accuracy. This process integrates spatial and temporal information into a higher-level representation, enabling the model to make informed decisions. The following point outlines the key aspects of our proposed work.

1. Context-Aware Preprocessing for Region of Interest (ROI) Identification: Our methodology uses a preprocessing phase with context-aware background subtraction to identify the ROI in the frames.
2. Two-Stream HAR Approach: Introduction of two-stream methodology, covering spatial and temporal feature extraction methods.
3. Late Binding Mechanism Architecture: The late binding mechanism is implemented to combine the outputs from the two streams.
4. Experimental Validation: Validation through experimentation to verify the effectiveness and dependability of the proposed methodology in various scenarios.

The paper addresses spatial and temporal feature extraction, late binding mechanism architecture, and rigorous experimental validation, highlighting the approach's reliability. Our contributions aim to bridge gaps in HAR research, establishing a robust foundation for precise activity recognition in real-world video data. The paper's format consists of a literature survey in Section 2, a discussion of the proposed system's approach in Section 3, and an experimental setup and result analysis in Section 4.

2 Literature Survey

Recognizing human activities through video analysis has become a pivotal focus within computer vision research. The literature survey explores diverse methodologies employed for the HAR system using video data, encompassing both deep learning approaches and those based on handcrafted features. With innovative approaches, like deep neural networks (like ResNet and CNNs) and creatively designed manual feature extraction procedures, researchers have significantly improved the accuracy and practicality of HAR systems. This survey provides a comprehensive overview of these approaches, underscoring their potential impact across various domains, from surveillance systems to elderly care. Human activity recognition using video data has been a significant area of research. Traditional human activity recognition started by extracting handcrafted engineered features and using off-the-shelf classifier algorithms. Numerous

articles have examined the HAR utilizing well-engineered features. In [6], authors presented a novel approach to HAR in their work, which generates feature vectors from compression variables through video coding. Each variable's statistical parameters across all video frames are calculated, effectively removing the temporal domain from the feature vectors. In [9], authors describe a novel approach to feature extraction that utilizes unsupervised feature learning methods such as PCA, denoising auto-encoder, and sparse auto-encoder to extract useful feature representations from the sensor data. An effective overlapping multi-feature descriptor and classification system for recognizing human activity is introduced by Cho and Byun in [4]. It captures both local and temporal information. In [5], the authors suggest a skeleton-based feature named Orientation Invariant Skeleton Feature (OISF) and employ it to train a Random Forest classifier for human activity recognition. In [12], the velocity histories of vital points are tracked to create features for activity recognition, which performs well on the KTH dataset and a new dataset of activities of daily living (URADL). As we delve into the advancements in human activity recognition, it becomes evident that traditional approaches have paved the way for a new era marked by integrating deep learning models. The subsequent discussion will focus on the challenges and breakthroughs introduced by Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and other innovative deep learning architectures in recognizing human activities from video data. Human activity recognition using video and deep learning has been complex in computer vision. Various approaches have been proposed to address this task. Deep learning models such as CNNs and RNNs have shown encouraging results. CNN-based fusion architectures like Long Term Recurrent Convolutional Networks (LRCN) and a combination of convolutional and Long Short-Term Memory (LSTM) layers called ConvLSTM, are applied with encouraging results [7][14][16]. These models have been applied to different datasets and achieved high accuracy in recognizing human actions and gestures [17]. Additionally, using pre-trained CNNs like ResNet50, ResNet101, InceptionV3, and InceptionResNetV2 has shown promising results in human activity recognition, achieving high accuracy on the Stanford 40 dataset [13]. There are diverse approaches to addressing video-related HAR tasks. One strategy involves leveraging deep learning models like ResNet and vision transformer architectures like ViT [11][14]. Additionally, researchers have explored incorporating motion information and utilizing feature extraction methods like Block-Based Motion Intensity Code (BBMIC) [1]. Activity recognition has also benefited from applying deep neural networks, including models like DenseNet121 [18] and MobileNetV2 [19]. Different benchmark datasets, such as HMDB51, UCF50, KTH, and Wizmann, have been used to evaluate the performance of these models. The literature survey highlights the diverse approaches in human activity recognition using video data. These approaches hold practical implications in domains such as surveillance, medical diagnostics, and employee activity monitoring, showcasing the significance of ongoing research in this field.

3 Methodology

In the proposed system, human activity is classified from the video data. In the initial step, all video frames go through the preprocessing phase, where the region of interest (ROI) is extracted from each frame of the video data. The proposed system achieves the ROI by performing context-aware background subtraction.

3.1 Context-Aware Background Subtraction

This preprocessing step involves applying a context-aware background subtraction technique to the video frames. Background subtraction is a method used to identify the foreground objects or regions in an image by subtracting the static background. The context awareness based on adaptive Gaussian Mixture Background Modeling is used, derived from the traditional Gaussian Mixture Models (GMMs) for background subtraction in computer vision. This approach incorporates contextual information to adapt the background model based on the scene's context. The Context-Awareness Based Adaptive Gaussian Mixture Background Modeling [20] relies on Gaussian Mixture Models (GMMs) as its foundation, utilizing a mixture of Gaussian distributions to represent pixel values in the background, with each Gaussian component capturing a distinct statistical property. The method incorporates Adaptation Over Time, similar to standard GMMs, adjusting its background model over time to address changes in the scene, including variations due to lighting, camera movement, or new object introductions. Context-awareness integration plays a crucial role in both Spatial Context and Temporal Context. Spatially, the model refines the background by analyzing relationships between neighboring pixels, and it is beneficial in scenarios with spatial dependencies among objects. Temporally, the model leverages contextual adaptation by considering the history of pixel values over time, distinguishing short-term anomalies from persistent changes. Weighted Contributions assign weights to Gaussian components based on their relevance in the current context, prioritizing more representative components and removing less relevant ones. Dynamic Learning Rates further enhance adaptability by adjusting update rates for model parameters based on the observed context, ensuring responsiveness to changes when needed and stability in stable scenes. Once the ROI is marked in the frames based on the context, the frames are fed into the spatial and temporal steams for further feature extraction and analysis.

3.2 Two Stream framework

Two-stream human activity recognition from video is an approach that utilizes two separate streams, each specializing in capturing different aspects of information from video data to improve recognition accuracy, as shown in Figure 1. These streams are typically called the "spatial stream" and the "temporal stream." An overview of how the proposed system is implemented is given below:

3.2.1 Spatial Stream

1. Feature Extraction: Utilizes CNNs and pre-trained models like ResNet to capture static visual details in video frames, such as body postures, object interactions, and scene context. The model is fine-tuned for spatial feature extraction.

2. Frame-Level Analysis: Independently processes each video frame in the spatial stream, focusing on extracting relevant features without considering the temporal context. Emphasizes what is visually present at a specific point in time.

3.2.2 Temporal Stream

1. Feature Extraction: Captures dynamic information about motion and action using a LSTM model and optical flow algorithm. Emphasizes temporal patterns by analyzing changes in visual features across frames.

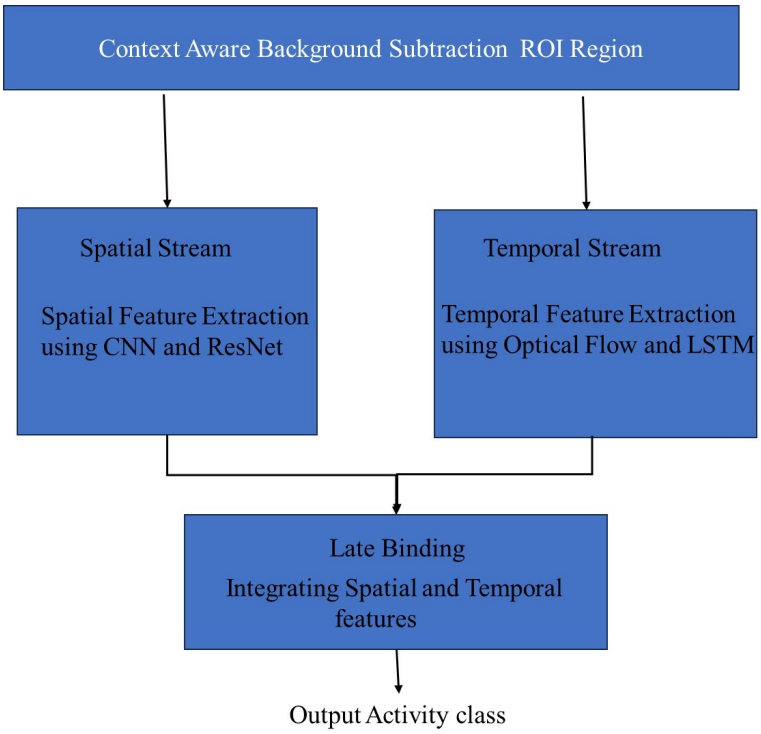

Figure 1: Block diagram of the proposed Two Stream System.

2. Sequence-Level Analysis: Considers the video's frame sequence to assess the evolution of features over time. It is crucial for understanding the progression of activities, unlike the spatial stream that analyses individual frames independently.

3.3 Late Binding Mechanism

Once spatial and temporal features have been independently extracted, the late binding mechanism combines them to provide a comprehensive representation for activity recognition. This integration is done by concatenating the two streams. The late binding mechanism enables the model to make informed decisions by exploiting the complementary nature of spatial and temporal information. The model can recognize activities more accurately by considering the scene's static and dynamic aspects. By incorporating both spatial and temporal streams, the two-stream approach enhances recognition accuracy and robustness, particularly in scenarios where understanding the visual context and motion dynamics is critical. It's an effective strategy for addressing the challenges associated with human activity recognition in video data.

4 Experiment and Result Analysis

4.1 Experimental Setup

4.1.1 Datasets

The experiment was conducted on two widely used video datasets: HMDB 51 (Human Motion Database) [8] and UCF (UCF50 Action Recognition Dataset) [15]. The UCF50 dataset is the largest publicly available action recognition dataset, with 50 action categories and 6676 videos. The HMDB51 dataset consists of realistic videos taken from YouTube. It contains 50 action categories.

4.1.2 Model Configuration and Evaluation

In our two-stream approach, combining spatial and temporal streams, we predict human activity. We evaluate performance by assessing individual stream performance on given datasets, evaluating the system with context-aware background subtraction for ROI, and comprehensively analyzing the overall model's performance with integrated spatial, temporal streams, and context-aware background subtraction.

4.2 Results Analysis

In Human Activity Recognition (HAR) using a two-stream approach, training loss minimizes differences in predicted and actual activities during training, while validation loss gauges model generalization. Training accuracy measures recognition within the training set, while validation accuracy indicates generalization ability. A notable gap between training and validation metrics suggests overfitting, requiring fine-tuning for robustness. Integration of spatial and temporal streams emphasizes validation accuracy's importance, crucial for real-world decision-making. Result analysis emphasizes these metrics, showcasing the high performance of the two-stream approach in HAR.

4.3 Spatial Model's Performance

The training history of the proposed spatial model on UCF datasets is represented using the graph, as shown in Figure 2. Our model demonstrates a noteworthy convergence as evidenced by the consistent decline in both training and validation losses across epochs.

4.4 Temporal Model's Performance

The training progress of the suggested temporal model on HMDB51 datasets is depicted through the graph illustrated in Figure 3, showcasing the loss and accuracy profile during the Training and Validation process.

4.5 Performance of the Combined Model

In this section, we present the training history of the proposed combined model on the UCF and HMDB51 datasets, which are represented in Figure 4 and Figure 5, respectively. The graphs in the Figure 4 and 5 show how the model performed as both training and validation losses decline per epoch and training and validation accuracy increases per epoch, which highlights effective learning, culminating in a significant improvement in the performance for HAR. The

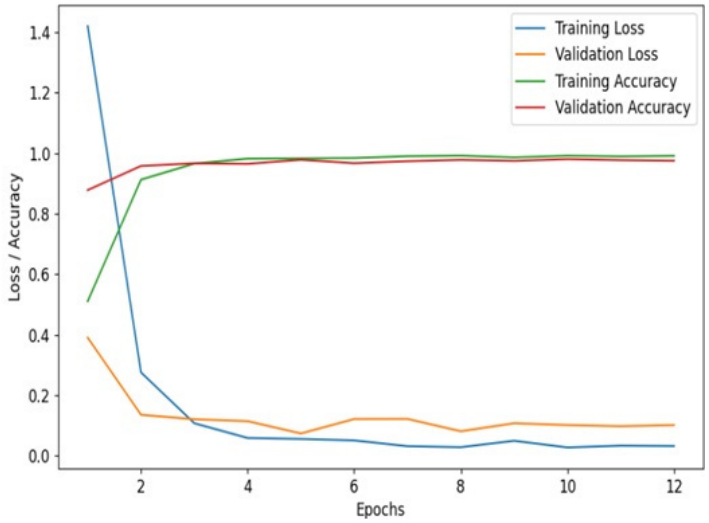

Figure 2: Training vs. validation history of Spatial model on UCF Dataset.

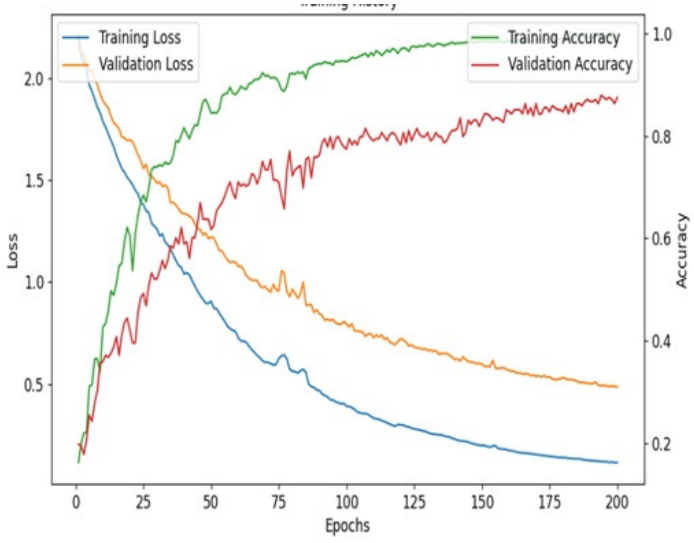

Figure 3: Training vs. validation history for the temporal model on HMDB51 Dataset.

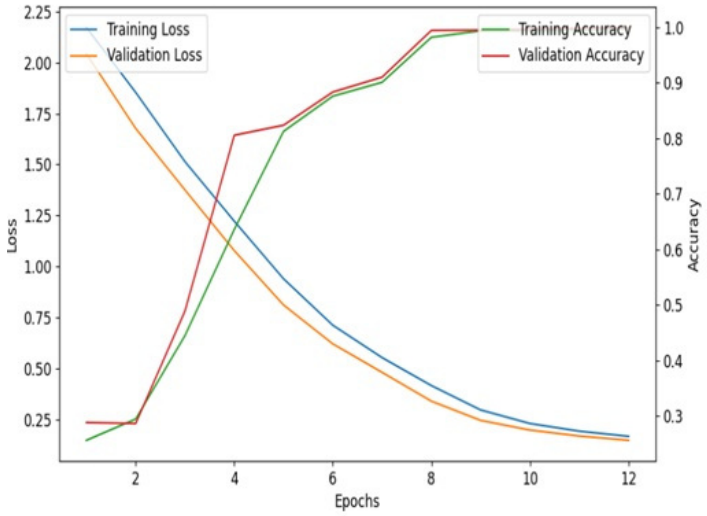

Figure 4: Training vs. Validation History of the Combined Model on UCF Dataset.

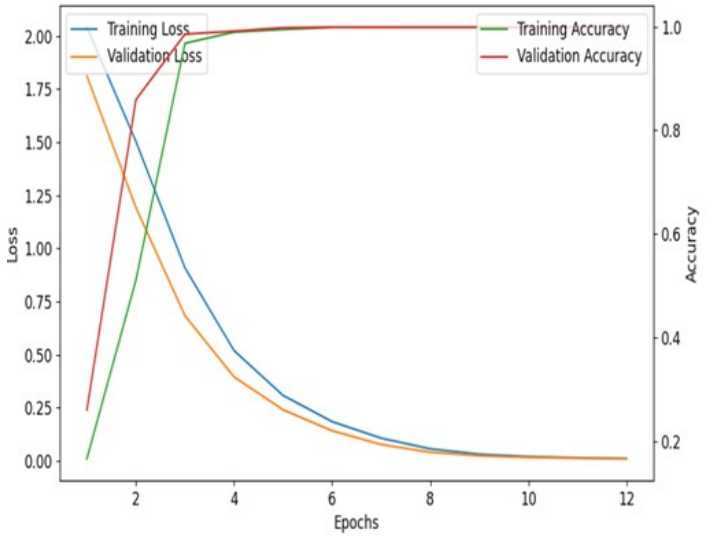

Figure 5: Training vs. Validation History of the Combined Model on HMDB 51 Dataset.

model's capacity to generalize is evident, as demonstrated by a high validation accuracy by the few epochs of training. These findings collectively emphasize the model's robust performance and its potential for reliable predictions on the UCF and HMDB51 Dataset.

4.6 Performance of the proposed system with context aware background subtraction

We used context-aware background subtraction as a preprocessing step to obtain the ROI. The effect of the context-aware background subtraction is shown in Table 1. The Context-Aware layer consistently improves the performance of all three models (Spatial, Temporal, and Combined) on both UCF and HMDB51 datasets, suggesting its effectiveness in enhancing the accuracy of the proposed system.

Model	Accuracy on UCF		Accuracy on HMDB51	
	Without	With	Without	With
Spatial Model	0.93	0.97	0.948	0.97
Temporal Model	0.72	0.78	0.81	0.87
Combined Model	0.95	0.99	0.952	0.99

Table 1: Performance of the two stream system with and without context aware background subtraction.

4.7 Evaluation of the Proposed Model

Table 2 summarizes the training and validation accuracies of three distinct models — spatial, temporal, and combined evaluated on both UCF and HMDB51 datasets. The Spatial Model exhibits high performance on both datasets, achieving 99.1% training accuracy and 97% validation accuracy on UCF, and 99.3% training accuracy and 96.8% validation accuracy on HMDB51. The Temporal Model shows slightly lower accuracies, particularly on the UCF dataset, with 90.8% training accuracy and 78% validation accuracy. In contrast, the Combined Model excels across the board, achieving 99.7% training accuracy and 99.7% validation accuracy on UCF, and an even higher 99.9% training and validation accuracy on HMDB51, showcasing its superior overall performance.

Model	UCF		HMDB 51	
	Training	Validation	Training	Validation
Spatial Model	0.991	0.97	0.9930	0.968
Temporal Model	0.908	0.78	0.994	0.875
Combined Model	0.997	0.997	0.999	0.999

Table 2: Training and validation accuracy on UCF and HMDB51 Datasets

Table 3 illustrates the performance of different models on the UCF and HMDB51 datasets. The Spatial Model achieves high precision, recall, and F1-Score values on the UCF dataset, exceeding 97%. These results indicate the model's proficiency in capturing static visual features and accurately recognizing activities. Conversely, the Temporal Model demonstrates slightly lower performance, with precision and recall around 85%, emphasizing its focus on dynamic information and motion patterns. However, the Combined Model, integrating spatial and temporal data, surpasses both individual models with precision, recall, and F1-Score values close to or exceeding 99%. This synergy highlights the model's effectiveness in human activity recognition (HAR) by leveraging both static and dynamic information. Similarly, on the HMDB51

dataset, the Spatial Model exhibits high precision, recall, and F1-Score, showcasing its ability to capture static visual features effectively. The Temporal Model, highlighting dynamic information, achieves slightly lower performance. The Combined Model, however, excels with precision, recall, and F1-Score values close to 99.9%. These results underscore the integration of spatial and temporal data in achieving superior HAR performance, particularly evident on the HMDB51 dataset. In summary, the Combined Model consistently outperforms individual Spatial and Temporal Models, emphasizing the synergistic benefits of integrating static and dynamic information for achieving superior recognition accuracy in HAR tasks.

Dataset	Model	Precision	Recall	F1-Score
UCF	Spatial Model	0.97	0.975	0.97
	Temporal Model	0.85	0.85	0.84
	Combined Model	0.99	0.99	0.99
HMDB51	Spatial Model	0.97	0.97	0.97
	Temporal Model	0.88	0.87	0.87
	Combined Model	0.99	0.99	0.99

Table 3: Evaluation and results on UCF and HMDB51 Datasets

Model	UCF Dataset	HMDB51
Flow-I3D	96.7	77.1
RGB-I3D	95.6	74.8
Two-Stream I3D	98	80.7
Two-Stream I3D, Kinetics pre-training	97.8	80.9
Flow-I3D	96.5	77.3
Proposed	99.7	99.9

Table 4: Comparison with the standard models [2]

From Table 4 The proposed system demonstrates high performance compared to existing models on the UCF and HMDB51 datasets. In the UCF dataset, the Proposed Model achieves an accuracy of 99.7%, surpassing other models, as shown in Fig 4. This points out the proposed approach's effectiveness in capturing both spatial and temporal information, significantly improving recognition accuracy. Similarly, on the HMDB51 dataset, the Proposed Model achieves an accuracy of 99.9%, outperforming all other models, as shown in Figure 5. The proposed system's high accuracy on both datasets demonstrates its robustness in handling diverse activities and its potential for real-world applications.The results highlight the significance of the proposed two-stream approach, which effectively integrates spatial and temporal information, leading to a highly accurate and versatile model for human activity recognition. The good performance across both UCF and HMDB51 datasets suggests the proposed system's efficacy in addressing challenges associated with recognizing human activities in video data.

5 Conclusion

Human Activity Recognition (HAR) from video signals is increasingly crucial for applications in surveillance, healthcare, robotics, and augmented reality. Accurately identifying human ac-

tions is vital in our data-driven world, posing a significant technological challenge. This study introduced a context aware two-stream framework, extracting spatial and temporal features in a complementary manner from the context-aware frames. The spatial stream analyses visual features from individual video frames, while the temporal stream focuses on dynamic aspects, capturing intricate motion patterns. In the next stage the late binding mechanism optimally allows interaction between spatial and temporal streams, improving recognition accuracy. The comprehensive experimental validation using HMDB51 and UCF50 datasets proves the proposed approach's efficacy and reliability. The result analysis showcased a high performance on the UCF and HMDB51 datasets, compared to the existing models. By analysing the individual spatial and temporal models, we observed the Spatial Model's precision in capturing static visual features within the HMDB51 dataset, while the Temporal Model exhibited slightly lower precision, emphasizing dynamic information. However, the Combined Model, integrating both streams, consistently outperformed with high precision, recall, and F1-Score values, showcasing the benefits of merging static and dynamic information. In summary, proposed system's high accuracy across diverse activities highlights its robustness and potential for real-world applications. This study not only contributes to advancing HAR technology but also provides a practical solution for computer systems to interpret and recognize human activities in videos, presenting a promising avenue for diverse real-world applications.

References

[1] Mariyan Richard A and Prasad N Hamsavath. Human Recognization Activity and Maximum Motion Representation in Surveillance Video. *EuropeanChemicalBulletin*, 12(2):2691–2697, 2023.

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[3] Wei Chang, Chunyang Ye, and Hui Zhou. Two-stream framework for activity recognition with 2d human pose estimation. In Aurélio Campilho, Fakhri Karray, and Zhou Wang, editors, *Image Analysis and Recognition*, pages 196–208, Cham, 2020. Springer International Publishing.

[4] SY Cho and HR Byun. Human activity recognition using overlapping multi-feature descriptor. *Electronics letters*, 47(23):1275–1277, 2011.

[5] Neelam Dwivedi, Dushyant Kumar Singh, and Dharmender Singh Kushwaha. Orientation invariant skeleton feature (OISF): a new feature for human activity recognition. *Multimedia Tools and Applications*, 79(29):21037–21072, August 2020.

[6] Obada Issa and Tamer Shanableh. Video-based recognition of human activity using novel feature extraction techniques. *Applied Sciences*, 13(11):6856, 2023.

[7] Yang Jiaxin, Wang Fang, and Yang Jieru. A review of action recognition based on convolutional neural network. In *Journal of Physics: Conference Series*, volume 1827, page 012138. IOP Publishing, 2021.

[8] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: A large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

[9] Yongmou Li, Dianxi Shi, Bo Ding, and Dongbo Liu. Unsupervised feature learning for human activity recognition using smartphone sensors. In *Mining Intelligence and Knowledge Exploration: Second International Conference, MIKE 2014, Cork, Ireland, December 10-12, 2014. Proceedings*, pages 99–107. Springer, 2014.

[10] Congcong Liu, Jie Ying, Haima Yang, Xing Hu, and Jin Liu. Improved human action recognition approach based on two-stream convolutional neural network model. *The Visual Computer*, 37(6):1327–1341, June 2021.

[11] Ravishankar Mehta, Sindhuja Shukla, Jitesh Pradhan, Koushlendra Kumar Singh, and Abhinav Kumar. A vision transformer-based automated human identification using ear biometrics. *Journal of Information Security and Applications*, 78:103599, 2023.

[12] Ross Messing. *Human activity recognition in video: extending statistical features across time, space and semantic context*. University of Rochester, 2011.

[13] Olena Pavliuk and Myroslav Mishchuk. A novel deep-learning model for human activity recognition based on continuous wavelet transform. In *IDDM*, pages 236–245, 2022.

[14] Hieu H. Pham, Louahdi Khoudour, Alain Crouzil, Pablo Zegers, and Sergio A. Velastin. Video-based human action recognition using deep learning: A review, 2022.

[15] Kishore K Reddy and Mubarak Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5):971–981, July 2013.

[16] Upasna Singh and Nihit Singhal. Exploiting video classification using deep learning models for human activity recognition. In Praveen Kumar Shukla, Krishna Pratap Singh, Ashish Kumar Tripathi, and Andries Engelbrecht, editors, *Computer Vision and Robotics*, pages 169–179, Singapore, 2023. Springer Nature Singapore.

[17] Guilherme Augusto Silva Surek, Laio Oriel Seman, Stefano Frizzo Stefenon, Viviana Cocco Mariani, and Leandro dos Santos Coelho. Video-based human activity recognition using deep learning approaches. *Sensors*, 23(14), 2023.

[18] Nusrat Tasnim, Mohammad Khairul Islam, and Joong-Hwan Baek. Deep learning based human activity recognition using spatio-temporal image formation of skeleton joints. *Applied Sciences*, 11(6), 2021.

[19] Zhou Xiaolong, Jin Tian, and Du Hao. A lightweight network model for human activity classifiction based on pre-trained mobilenetv2. 2021.

[20] HongGang Xie, JinSheng Xiao, and JunFeng Lei. Context-awareness based adaptive gaussian mixture background modeling. In Shin'ichi Satoh, editor, *Image and Video Technology*, pages 415–425, Cham, 2018. Springer International Publishing.