Journal of Engineering and Technology for Industrial Applications

ISSN ONLINE: 2447-0228

ITEGAM-JETIA

Manaus, v.12 n.58, p.1368-1376. March/April, 2026.
DOI: https://doi.org/10.5935/jetia.v12i58.3346

RESEARCH ARTICLE OPEN ACCESS

A MULTIMODAL GRAPH CONTRASTIVE LEARNING FOR HUMAN ACTIVITY RECOGNITION USING DEEP LEARNING TECHNIQUE

Velantina V[*1], V. Manikandan[2] and P. Manikandan[3]

[1]Research scholar, Department of Computer Science and Engineering, Jain (Deemed-to-be-University), Bengaluru, Karnataka, India.
[2]Assistant Professor, Department of Computer Science and Engineering, Jain (Deemed-to-be-University), Bengaluru, Karnataka, India.
[3]Professor, Department of Computer Science and Engineering, Jain (Deemed-to-be-University), Bengaluru, Karnataka, India.

[1]https://orcid.org/0009-0006-1905-3470, [2]https://orcid.org/0000-0002-8620-1860,[3]https://orcid.org/0000-0003-3037-7688

Email:*velantinav@gmail.com, v.manikandan@jainuniversity.ac.in, p.manikandan@jainuniversity.ac.in

ARTICLE INFO

Article History
Received: January 19, 2026
Reviewed: February 23, 2026
Accepted: April 2, 2026
Published: April 30, 2026

Keywords:
Human Activity Recognition,
Temporal Transformer Network,
Multimodal Fusion,
Transformer,
Deep learning,
Hybrid Optimization

ABSTRACT

In recent years, the deep learning technique for Human Activity Recognition (HAR) systems has made remarkable improvements in recognising complicated activity classes and real-world scenarios. This study introduces a unified deep learning framework, called the Hybrid Dense Temporal Transformer Network (HDTTN) is used to capture spatial, temporal, and semantic information for improved human activity detection. Using the DenseNet-201 to improve spatial feature extraction from visual inputs, Temporal Convolutional Networks (TCNs) for learning short-term motion patterns, and Transformer encoders for learning long-range temporal dependencies that are crucial for processing complex and subtle activities. This study employs an early multimodal feature fusion strategy to further enhance representational coherence, which makes it easier to incorporate heterogeneous cues at the feature level and to learn dynamic multimodal representations. Moreover, a hybrid optimization approach is integrated for parameter fine-tuning for efficiency, reduced overfitting, and increased model robustness. The proposed HDTTN framework is shown to be effective on the large-scale Kinetics dataset containing a wide range of unconstrained human activities. Experimental results shows that the proposed model is 93% accurate, compared to several existing state-of-the-art baseline approaches. Moreover, qualitative and quantitative analyses validate HDTTN's ability to identify intricate and complex activities across a multitude of environments.

I. INTRODUCTION

Human Activity Recognition (HAR) has become a major research area in computer vision and artificial intelligence, which refers to the task of identifying and understanding human actions in video data automatically. The proliferation of surveillance cameras, digital media platforms, and intelligent video analytics systems has led to the continuous generation of massive visual data in real time, which requires automatic approaches that are able to interpret complex human behaviors within everyday life settings [1]. Thus, video-based HAR has drawn much attention due to its important applications in public safety surveillance, health care video monitoring, sports performance analysis, video retrieval, and human–computer interaction [2]. Video-based HAR is learning spatiotemporal representations from RGB videos, depth sequences, optical flow, or pose-based representations.

Compared with still image recognition, action recognition needs to jointly model the spatial appearance, temporal motion dynamics, and changes of context over time, which is a more challenging learning problem [3]. Dynamics may vary from simple daily behaviour's to complex activities with long-term temporal dependencies and subtle motion changes. Therefore, HAR systems must consider both short-term motion cues and long-range temporal dependencies [4]. Early video-based HAR methods were based on manually designed primitives with spatiotemporal interest points, motion descriptors with optical flow and trajectory-based representations, along with traditional classifiers. Although these methods showed an acceptable performance in constrained situations, they had a strong response to background clutter, viewpoint difference (VPD), and light change and were restricted to practical use [5].

The era of deep learning significantly advanced video-based HAR, where discriminative representations can be learned in an end-to-end manner directly from raw video data. Summarization regarding the spatial feature extraction, CNNs are widely considered to be effective at learning hierarchical visual representations. Two-stream CNN architectures additionally enhanced the recognition performance by encoding appearance and motion information jointly with RGB frames alongside optical flow [6]. Then, 3D-CNNs generalized the convolutions in a three-dimensional space and learned the spatial-temporal characteristics simultaneously [7]. However, 3D CNNs are computationally expensive and have inherent difficulties in modelling long-range dependencies because they are based on fixed-size temporal receptive fields.

Sequence-based models are employed to address the limitation of exclusively relying on frame information (e.g., Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have been proposed to learn temporal dependencies among video frames [8]). However, such models can have vanishing gradients and are not scalable to long videos. Temporal Convolutional Networks (TCNs) have been proposed as an efficient alternative that provides gradient stability and parallel computation in modelling short-term and long-range temporal dependencies [9]. However, TCNs may still be restricted in their ability to encapsulate the intricate long-range dependencies that are present in large-scale human activities independently. The graph-based HAR approaches provide more attention because they can model the local structured relationship between body joints, regions, or pixels over time explicitly.

In these works, videos are modelled as graphs in which nodes represent spatial entities and edges denote either spatial or temporal interdependences [10]. Graph Convolutional Networks (GCNs) effectively extract non-Euclidean dependencies and correlation structures that are extremely difficult to represent using classical convolutional operations. By utilizing graphs, such approaches improve on explicitly modelling spatiotemporal interaction relationships that are necessary for activity recognition. Recent research has expanded temporal graph modelling and contrastive learning to new levels of sophistication to enhance the generalization and robustness of HAR. Temporal graph contrastive learning (TGCL) acts to align positive temporal representation pairs and suppress out-of-class motion, which has the potential of reducing intra-class variance and promoting discrimination [11].

The performance of the majority of existing graph-based methods can be compromised by complex activities, as they only consider limited temporal context or do not integrate mechanisms to incorporate global temporal dependencies, despite being driven by their effective results. The inadequate generalization across datasets and environments is the main cause of the reason HAR on video remains an unsolved problem. It has been stated that the distribution shift between training and testing environments (screened camera view, background, actor face, and movement style) can lead to a significant change in appearance [12]. Additionally, deep HAR models are highly susceptible to network architectures and hyperparameter selection, which could result in models that are unstable and less robust if not methodically calibrated.

Inspired by these challenges, in this paper we propose a Hybrid Dense Temporal Transformer Network (HDTTN) for the task of robust video-based HAR. The proposed framework combines DenseNet-based spatial feature extraction to improve feature reusing and gradient propagation, Temporal Convolutional Networks (TCNs) to describe the short-term motion dynamics, and Transformer encoders for learning long-range temporal dependencies with self-attention mechanisms. In addition, graph-based temporal modelling methods are included to efficiently model spatiotemporal relationships. An advanced feature-level fusion approach and a hybrid optimization scheme are used to increase the generalization power while preventing overfitting, as well as to improve recognition performance. The objectives of this study consist of:

- To construct a unified HDTTN for video-wise human action understanding.
- To achieve an effective early feature fusion method and graph-informed temporal reasoning.
- To optimize the HDTTN model with hybrid optimization methods that improve the efficiency of fine-tuning all parameters.

II. LITERATURE REVIEW

Human activity recognition (HAR) with RGB video is a significant issue with possible applications in surveillance, healthcare monitoring, sports analytics, and intelligent human-computer interaction systems [13]. Recent developments in deep learning techniques have enhanced the capability of learning spatiotemporal representations directly from raw video streams without the need for predefined descriptors. The most recent studies emphasize the importance of convolutional and attention-based architectures in video-based HAR research [14]. They also highlight challenges with long-range temporal modeling and robustness to complex real-world conditions.

Initially, deep approaches relied on 2D and 3D Convolutional Neural Networks (CNN) to detect spatial appearance and short-term motion [15]. 3D CNNs are effective for learning local spatiotemporal characteristics but have high computational costs and narrow temporal receptive windows [16]. Generalized two-stream CNNs with temporal integration were effective for short and basic activities but lacked global temporal reasoning for long and overlapping tasks. To solve temporal restrictions, a combination of CNN and RNN was developed. The CNN-LSTM cascades to encode frame-level characteristics and sequential information [17].

These techniques improved temporal continuity but exhibited limitations in sequential processing and capturing long-range interdependence in complex operations [18]. The development of Temporal Convolutional Networks (TCNs) represents a shift toward parallel temporal modelling. Temporal convolutions with dilation effectively capture multi-scale motion information [19], [20]. These efficient models focused on temporal dynamics and lacked global contextual reasoning for video content [21]. The HAR model using the Vision Transformer, known as NiViT, demonstrated significant improvements over models that rely solely on CNNs [22].

Transformer-based models require a huge amount of pre-trained data and expensive computation, making them unsuitable for practical applications. Hybrid CNN-Transformer architecture's efforts to balance efficiency and accuracy have yielded high recognition accuracy on large-scale benchmark datasets [23], [24]. Graph-based HAR approaches have improved video interpretation by directly displaying related structures. GCNs were used to model inter-frame (or region-level) dependencies, increasing resilience against occlusions and perspective alterations [25]. However, graph-based models primarily capture poses and limited temporal information, lacking dense appearance modelling [26].

Recently, a number of multimodal video-centric methods have been developed that integrate RGB appearance with semantic information and activity. Alternatively, they implemented early and late fusion techniques to enhance their discrimination capabilities. Even though the recognition capability was improved by the fusion, the scalability was not possible due to the suboptimal feature alignment and inadequate temporal smoothness [27], [28]. Comparative studies have shown that the majority of current video-based HAR models achieve an accuracy of 85–92%, particularly when tested on challenging datasets such as Kinetics. Optimization-driven HAR models have also been considered to enhance learned knowledge, and they employ evolutionary and meta-heuristic optimization for the fine-tuning of deep networks [29].

The improvement was feasible however, it was unsuccessful in addressing the instability of convergence and the complexity of training [30]. Despite these advancements, video-based and graph-based HAR approaches still face three major obstacles, including their inability to reliably integrate spatial, temporal, and semantic information, and they do not adequately capture long-term spatio-temporal correlations [31], [32]. These gaps inspire the designs of the proposed HDTTN, which combines transformer-based global attention with optimized early feature fusion, TCN-driven temporal modelling, and Dense Net based spatial encoding by overcoming the drawbacks of earlier research and providing superior recognition performance.

From the extensive literature review described, we can see that the current video-based and graph-based HAR models either focus on spatial appearance or short-term temporal dynamics or long-range attention. CNN based methods are not able to capture long-range temporal dependencies, and transformer-based models are effective at representing global context but computationally inscrutable and inefficient with limited spatial capability. Graph-based models advance relational reasoning, yet mostly based on abstract representations which ignore dense visual semantics. Moreover, traditional fusion approaches are not efficient enough and suffer from the shattered feature representation, whose scalability is inefficient.

Inspired by these limitations, we propose a Hybrid Dense Temporal Transformer Network (HDTTN) that presents an end-to-end architectural solution to fill out the above-mentioned technical voids in a synergistic approach through densely connected spatial feature propagation with DenseNet, fine-grained motion modeling with Temporal Convolutional Networks, and long-range temporal reasoning by means of transformer encoders. Early fusion strategy with optimization over the multimodal space yields not only consistent but also temporal coherent representations, and hybrid objective improves the generalization and robustness. This end to end design could directly overcome the limitation associated with prior arts and build up a scalable, high-accuracy framework for real-time human activity recognition.

It is apparent from the comprehensive literature review that the current HAR models, which are based on graphs and videos, either concentrate on spatial appearance, short-term temporal dynamics, or long-range attention. CNN-based methods are incapable of capturing long-range temporal dependencies, while transformer-based models are effective in representing global context but computationally perplexing and inefficient, with limited spatial refining capabilities. Graph-based models make significant contributions to relational reasoning however, they are primarily based on abstract representations that disregard complex visual semantics. Moreover, conventional fusion methods are insufficiently efficient and are complicated by the fragmented feature representation, which has poor scalability.

The proposed Hybrid Dense Temporal Transformer Network (HDTTN) is a synergistic solution to address the aforementioned research gaps. This entire design solution comprises densely connected spatial feature propagation with DenseNet, fine-grained motion modelling with Temporal Convolutional Networks, and long-range temporal reasoning through transformer encoders. This strategy is driven by these constraints. Early fusion methods that optimize over the multimodal space produce representations that are consistent and temporally coherent, while hybrid methods enhance the generalization and robustness. This end to end design could directly overcome the limitation associated with prior arts and build a scalable, high-accuracy framework for real-time human activity recognition.

II.1 RESEARCH GAP

Despite the remarkable advancements in video-based and graph-based Human Activity Recognition (HAR), several significant research gaps remain insufficiently addressed in recent years.

- **Insufficient Long-Term Temporal Modeling:** While most methods based on CNN or TCN focus on short-term motion patterns, none of them model long-range temporal dependencies required for recognizing complex and composite activities.
- **Fragmented Feature Fusion Strategies:** Current hybrid networks often perform late fusion or loosely coupled feature integration, which can cause semantic divergence and loss of information from the spatial, temporal, and relational signals.
- **Limited Exploitation of Graph-Based Semantics:** Though graph-based representation is widely adopted, most methods rely on quality of pose estimation or shallow graph reasoning, which limits robustness in unconstrained real-world videos.
- **Scalability and Generalization Issues:** Language-only or 3D CNN models require significant pretraining and are computationally intensive to both real-time deployment and generalization across datasets.

III. PROPOSED METHODOLOGY

The proposed Hybrid Dense Temporal Transformer Network (HDTTN) model aims to coordinate multimodal data input from video, audio, and text for effective human activity recognition. The approach is based on four major steps, including multimodal preprocessing, feature extraction, early fusion, and classification using a hierarchical deep temporal transformer network modified by dual-phase wasp monkey optimization. This study focuses on the HDTTN model which is used to predict human activity by integrating multi-modal data (video, audio, and text) through the use of various methods that synergize dense temporal modelling and multimodal fusion, optimized via Wasp Optimization for enhanced accuracy, scalability, and robustness across complex HAR tasks. It encompasses a variety of phases, including data preprocessing, feature extraction, multimodal fusion, classification, and optimization, as illustrated in Figure 1.

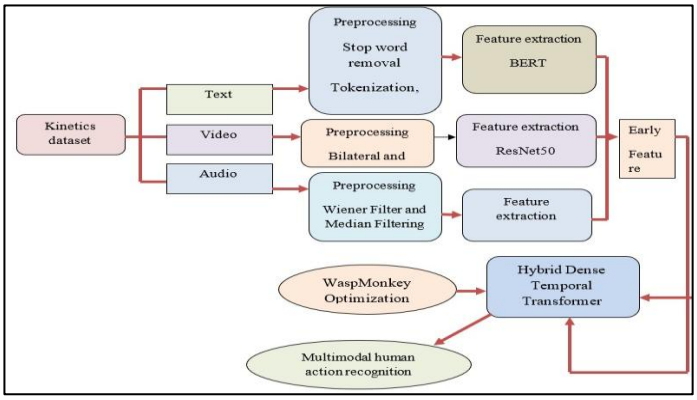

Figure 1: Proposed Multi-Modal Framework for HDTTN.
Source: Authors, (2026).

The above Figure 1 represents the architecture diagram, which describes the steps involved in activity recognition. It contains thousands of video clips, each containing marked human actions, allowing it to be a large dataset. Each video clip is marked with an action and contains video and audio, with the possibility of additional audio text, captions, or subtitles, allowing the dataset to be leveraged for multimodal learning.

- Pre-processing: Video frames are pre-processed with bilateral filtering and CLAHE, audio signals are subject to denoising by wiener and median filters, and text is normalized by applying stop-word removal, tokenization, and lemmatization.
- Feature Extraction: The spatial video features are extracted by ResNet50, the acoustic features are extracted by MFCC, and the contextual text features are extracted by BERT.
- Early Fusion: Vector features are concatenated into a single global representation.
- Hierarchical Deep Temporal Transformer Network (H-DTTN): Dense Net enables feature enrichment, TCN learns temporal dependencies, and Transformer layers capture long-range global interactions.
- Optimization: The model's parameters are modified by Wasp Optimization, Dual-phase of optimization, WSO performs local refinement, while SMO explores global solutions.

III.1 DATA COLLECTION

The input source video is collected from the Kinetics dataset. It is a subset of the Kinetics-400 and Kinetics-600 datasets for video action recognition. It seeks to offer a representative and systematic sample for the purposes of training and evaluation. The subset contains over 10,000 short video clips, each about 10 seconds long. It contains 10 to 40 samples for each class of a wide range of human actions. The videos are sourced from YouTube, and as a result, they are diverse in the backgrounds, lighting, motion, and camera angles, which are ideal for challenging human action recognition problems. Each clip is linked to a singular human action and accompanied by video, audio, and other multimodal data to enable multimodal learning, as shown in Equation 1.

$$d_i = (V_i, A_i, T_i), \forall i \in [1, N] \quad (1)$$

Where d_i represents a single sample from dataset, N the total number of samples in the dataset, V_i is the raw video data for sample with dimension $T \times H \times W \times 3$, A_i is the raw audio signal for sample i, T_i is the raw text data for sample i.

III.2 PRE PROCESSING

Text Modality: It is a vital component of the multimodal framework, providing semantic context that may not be apparent from visual or aural streams alone. Textual information, including speech, video captions, and scene descriptions, is frequently used to describe or accompany human actions. This modality assists in the disambiguation of actions that may appear visually or acoustically similar but differ in intention, instruction, or dialogue. Textual data is pre processed for instance, "the" or "is" are frequently occurring words with little meaning that will be removed through stop word removal. Prepares the text for embedding by splitting the text into words or sub-words through a process called tokenization, as shown in Equation 2.

$$T_i^{(2)} = S(T_i) \quad (2)$$

Audio modality: It plays a critical role in capturing complementary information that is often not visually observable, such as speech, environmental sounds, or background cues.

To ensure the audio data is clean, clear, and suitable for feature extraction, two key preprocessing techniques, Wiener filtering and median filtering, are applied sequentially. Wiener filtering is a statistical signal processing technique used to reduce background noise in an audio signal. Processing of audio is represented in Equation 3.

$$A_i^{Wiener}(t) = \frac{S_{xx}(t)}{S_{xx}(t)+S_{nn}(t)} . A_i(t) \quad (3)$$

Here, k is the window size of the median filtering, A_i^{Wiener} and is the median filtered signal of time t.

Video Modality: Preserving action-related details while removing noise is accomplished by smoothing techniques like bilateral filtering in the video frame, which shows the bilateral filtered frame as neighbouring pixels surrounding the pixel using the spatial Gaussian function BF. The bilateral range Gaussian function is represented by the intensities G_s and G_r, which serve as the normalizing factor. In Equation. 1 represents the bilateral filtering with its frames and normalizing factor by eliminating the noise in the video as shown in Equation 4.

$$f_{ij}^{BF} = \frac{1}{W_p}\sum_{q\in\Omega} G_s\left(\|p-q\|\right)G_r\left(\left|f_{ij}(p)-f_{ij}(q)\right|.f_{ij}(q)\right) \tag{4}$$

Here, f_{ij} indicates the j^{th} frame of the i^{th} video, f_{ij}^{BF} is the Bilateral-filtered frame, Ω is the neighborhood pixels around the pixel p, the spatial Gaussian function is Gs, the range Gaussian function intensity is G_r and W_p is the normalizing factor.

III.3 FEATURE EXTRACTION

Text: Feature extraction using BERT plays a vital role in the text modality of this multimodal human action recognition framework. After completing the NLP preprocessing steps like stop word removal, tokenization, and lemmatization, the cleaned and structured text is fed into the BERT model to generate rich, context-aware embeddings, as shown in Equation 5.

$$T_i = \varphi_t\left(T_i^{clean}\right) \in \Re^{d_t} \tag{5}$$

Here, φ is the BERT transformer model, T_i is the extracted text feature vector and d_t is the text embedding size.

Audio: After denoising the audio signals using Wiener and median filtering, MFCCs are extracted to represent the audio in a way that aligns with human auditory perception. The MFCC algorithm transforms the audio signal from the time domain to the frequency domain, as denoted in Equation 6.

$$A_i = \varphi_a\left(A_i^{Med}\right)DCT\left(log\left(M.\left|FFT\left(A_i^{Med}\right)\right|^2\right)\right) \tag{6}$$

Here, φ_a represents the MFCC transform function A_i representing the extracted audio feature vector, FFT is short for Fast Fourier Transform, M is short for Mel-filter bank matrix, DCT is short for Discrete Cosine Transform, and d_a represents the audio feature size.

Video: The ResNet50, a deep CNN with 50 layers, is employed in this research for spatial feature extraction from enhanced video frames. After applying bilateral filtering and CLAHE to improve frame quality and contrast, each frame is fed into the pretrained ResNet50 model, which is known for its robust residual learning framework, as shown in Equation 7.

$$V_i = \varphi_v(V_i^{CLAHE}) \in \Re^{d_v} \tag{7}$$

Here, φ_v denotes ResNet50 model for spatial feature extraction, V_i is the extracted spatial feature vector for a video data and d_v represented by feature dimension.

III.4 EARLY FUSION OF MULTIMODAL FEATURES

After extracting modality-specific features from video using ResNet50, audio using MFCC, and text using BERT, the information from these three modalities is combined through an early fusion process. This involves concatenating the individual features into a single fused multimodal feature vector. For both the Kinetics datasets, this fused representation captures comprehensive spatiotemporal, acoustic, and semantic cues related to the human actions depicted in each video sample. These fused feature vectors serve as enriched input for the next stage of processing within the HDTTN, enabling the model to learn cross-modal correlations and temporal dynamics more effectively for improved action recognition accuracy, as represented in Equation 8.

$$F_iKi = concat(V_i, A_i, T_i) \in \Re^{d_v+d_a+d_t} \tag{8}$$

Here, F_iKi denotes the fused input vectors for Kinetics dataset.

III.5 HYBRID FUSION AND CLASSIFICATION

In the HDTTN, the final stage involves the hybrid fusion of outputs from the three parallel processing modules: the Dense Net block (spatial features), the Temporal Convolutional Network (temporal dependencies), and the Transformer encoder (global context). After each module has processed the same fused multimodal input, independently extracting features specific to its intended strength, these three outputs are concatenated to form a single, unified hybrid feature representation, as shown in Equation 9 below.

$$Z_{hybrid} = \text{concat}(z_{dense}, z_{tcn}, z_{trans}) \in \Re^{d_{hybrid}} \tag{9}$$

Where, $d_{hybrid} = d_{dense} + d_{tcn} + d_{trans}$, the final hybrid feature vector combining outputs from DenseNet, TCN, Transformer are denoted as Z_{hybrid}, and $concat(.)$ for the three modules, d_{hybrid} indicates total dimension by hybrid fusion as $d_{dense} + d_{tcn} + d_{trans}$.

III.6 WASP MONKEY OPTIMIZATION ALGORITHM

Wasp Monkey Optimization: To enhance parameter tuning and learning efficiency in the HDTTN model for multimodal human action recognition, the WMO algorithm is introduced as a hybrid metaheuristic. It combines the strengths of two bio-inspired strategies: WSO and SMO. WSO is employed for local exploitation, focusing on refining promising solutions by closely exploring their neighborhood to fine-tune model parameters such as weights and biases. In the final classification stage of the HDTTN, the hybrid feature representation obtained from the fusion of Dense Net, TCN, and Transformer outputs is passed through a fully connected classification layer. This layer transforms the high-dimensional hybrid feature vector into a set of predicted class probabilities, each representing the likelihood that the input video instance belongs to a specific human action category, as shown in Equation 10.

$$\hat{y}_i = soft\,max\left(W_c.Z_{hybrid} + b_c\right) \tag{10}$$

Here, $\hat{y}_i$ denotes the predicted class probabilities, W_c denotes the weight matrix of classifier and b_c denotes the bias vector and $soft\,max(.)$ denotes the softmax function to convert logits into probabilities.

IV. RESULTS AND DISCUSSION

In this section, provide an extensive experimental evaluation and analytical discussion for the proposed HDTTN-WMO framework. The main focus of this study is to analyse and compare the effectiveness of multi-modal feature fusion, utilizing a hierarchical temporal approach in conjunction with a dual-phase optimization strategy, which could be used to improve classification performance and generalization capabilities on challenging human activity categories involving wide varieties. The performance of the proposed model is systematically evaluated on a representative subset of the Kinetics dataset and compared with state-of-the-art baseline methods.

IV.1 DATASET DESCRIPTION

The experiments were done using a 5% subset of both Kinetics-400 and Kinetics-600, covering more than 10,000 short videos each with an average length of around 10 seconds. The dataset consists of a diverse set of human activities, with 10–40 samples per class that are to be normalized for both class diversity and balanced coverage. Videos were collected from YouTube, and the contents are diverse in terms of background scene, lighting condition, camera angle, and motion patterns, making it a challenging test for real-world HAR systems showing the effectiveness and efficiency of the approach with respect to moderate resource configurations.

IV.2 EXPERIMENTAL SETUP

In the context of this study, 70% of the data was trained, and the remaining 30% was subjected to testing. The study was conducted on a 2.4 GHz Intel(R) processor with 16 gigabytes of random-access memory (RAM) were utilized.

IV.3 EVALUATION METRICS

In order to obtain a comprehensive performance evaluation, multiple metrics were used, such as accuracy, precision, recall, and F1-score. Accuracy represents the correctness of classification overall, while precision and recall measure the reliability and completeness of predicted action classes, respectively. The F1 score gives an equal weight for both precision and recall. Confusion matrix analysis was further performed to obtain class resolution and to establish misclassification trends. The F1 score has been calculated as shown in Table 1 and Figure 2.

IV.4 CONFUSION MATRIX HEATMAP

There are several reasons that may explain substantial improvement. First and foremost, since the audio, video, and textual feature information are combined together earlier, the model is able to take advantage of complementary information across multiple modalities, which alleviates confusion for visually similar actions. Second, the hierarchical architecture of HDTTN, combining Dense Net, TCN, and Transformer blocks, is conducive to capturing spatial representations and temporal dependencies across multiple scales. Third, based on the dual-phase Wasp optimization, it can optimally tune hyperparameters, which would result in better convergence and generalization performance.

The confusion matrix is shown in Figure 2 on the Kinetics dataset that is very close to diagonal dominant, i.e., the action classes are accurately assigned in most cases. Very few misclassifications can be found, most of them between the action with a very similar temporal structure to the one imitated or in contextually related actions. Such behaviour and the proposed multimodal framework demonstrate how subtle temporal and semantic differences between similar classes can be learned. The generalization efficacy demonstrated by the confusion matrix confirms that the proposed optimal fusion strategy and hierarchical temporal mapping have been effective in dealing with the variation and complexity present in real human actions.

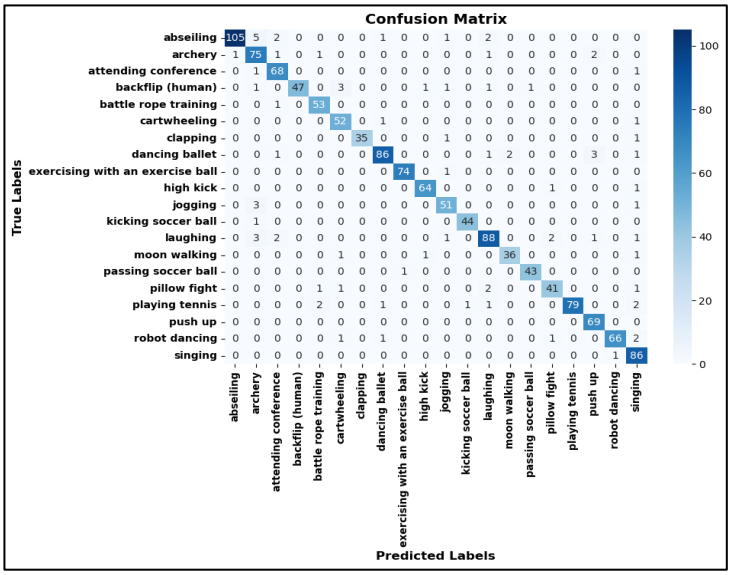

Figure 2: Confusion matrix showing classification results for Kinetics action classes.
Source: Authors, (2026).

Results from the Kinetics dataset show the flexibility of the proposed model with different levels of human motion complexity. The dataset's multimodal video, audio, and text components provided an opportunity to analyse the framework's scalability, resilience, and complexity. The results verified that the proposed model is robust and thus is very suitable for real-world human activity recognition systems.

Table 1: Accuracy Comparison.

Model	Accuracy (Kinetics)
CNN+LSTM	81.60%
2D-CNN	84.50%
VGG11	80.70%
AlexNet	86.50%
Proposed	93.00%

Source: Authors, (2026).

The model's per-class precision, recall, and F1-score performance are illustrated in Figure 3. The majority of the activity categories exhibit robust mean values, with a mean value exceeding 93%. Scores are slightly lower for actions that involve complex motion or rapid movement, which may be attributed to the interference of temporal spatial patterns. However, the system's stability is sufficient to accommodate a variety of motion types, as indicated by the combination of precision and recall. As shown in Table 1, the proposed multimodal HAR model outperforms the conventional models CNN+LSTM, 2D-CNN, VGG11, and AlexNet, all of which have lower accuracy. The results conclusively establish the higher classifier's performance, which is reliant on multimodal inputs and its adaptability across various video datasets. The hybrid model's consistent performance can be attributed to the remarkable synergistic impact of the DenseNet, TCN, Transformer layers, and the model's hyperparameters using whale optimization.

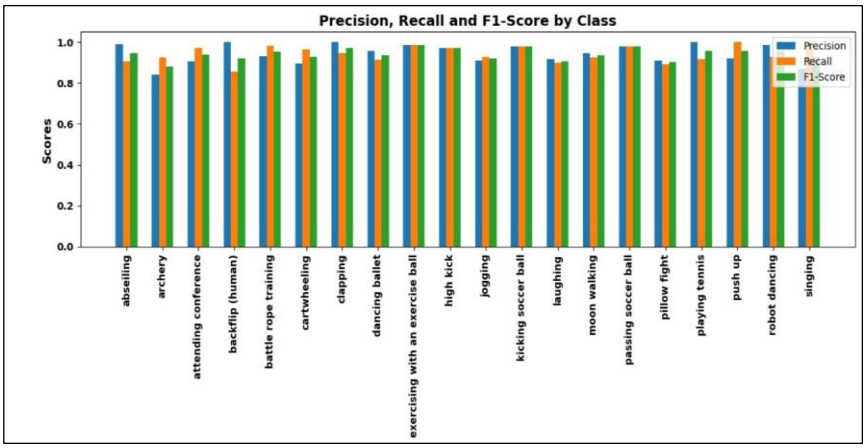

Figure 3: Precision, Recall and F1-Score across various activities.
Source: Authors, (2026).

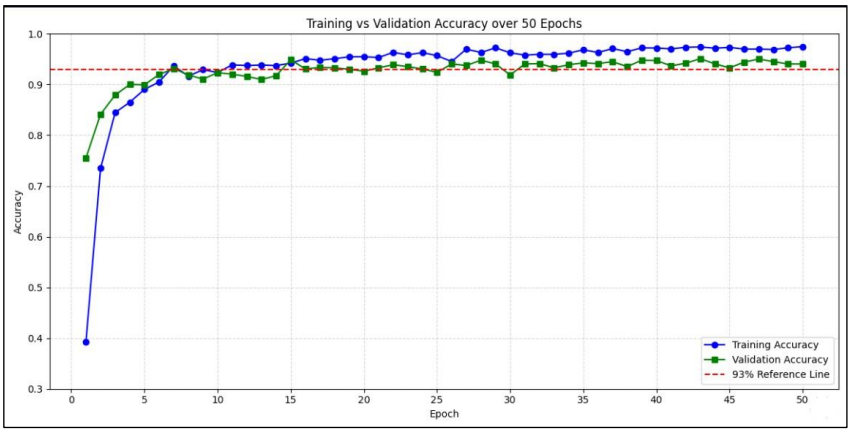

Figure 4: Comparison graph with various Baseline Models.
Source: Authors, (2026).

In Figure 4, the precision, recall, and F1-score of the HAR model of various actions performed by humans are represented. The model's performance and generalization when trained on the Kinetics dataset are represented by the accuracy and loss metrics evaluated over 50 epochs. The proposed approach achieves faster convergence, lower loss values, and higher peak accuracy, which shows that the learning is more efficient with less overfitting. Compared to the baseline model, the addition of WMO is important for guidance in the learning process and to create a balance between exploration and exploitation during the optimization of parameters.

V. CONCLUSION

In the context of multimodal HAR, this study proposed a robust and scalable HDTTN framework that simultaneously integrates video, audio, and textual information using HDTT models through dual-phase metaheuristic optimization. This approach outperforms traditional deep learning models in terms of accuracy, precision, recall, and F1-score, as indicated by comprehensive experimental results on the Kinetics dataset. The experimental results suggest that the synergistic interaction of Dense Net, Temporal Convolutional Networks (TCNs), Transformers, and Wasp-Monkey Optimization can effectively exploit complex spatiotemporal and semantic patterns with a high degree of generalization. In spite of these encouraging findings, there are still numerous prospective research avenues that demand further exploration. In order to eliminate dependence on labelled data, it would be beneficial to incorporate contrastive and self-supervised learning into future research. The discussion of small architectures and the interference of models with their base structure's information will be emphasized in the context of real-time inference on peripheral devices and those with limited resources.

Additionally, the proposed framework's application to related tasks (e.g., social interaction analysis and multimodal emotion recognition) and its extension to cross-domain dataset evaluation would be intriguing next steps. The HDTTN has demonstrated the highest level of efficacy with 93% in multimodal HAR, which integrates text, audio, and video modalities. Two-phase WMO optimization and HTM significantly enhance accuracy, generalization, and robustness. Future work will encompass the real-time integration of peripheral deployment for surveillance and healthcare applications, along with contrastive learning and fine-tuning. Enhances multimodal interaction and activity recognition, while improving human activity recognition in numerous real-world scenarios.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Velantina V, V. Manikandan, P. Manikandan.
Methodology: Velantina V, V. Manikandan.
Investigation: Velantina V, V. Manikandan, P. Manikandan.
Discussion of results: Velantina V, V. Manikandan, P. Manikandan.
Writing –Original Draft: Velantina V.
Writing –Review and Editing: Velantina V, V. Manikandan, P. Manikandan.
Resources: Velantina V, V. Manikandan, P. Manikandan.
Supervision: Velantina V, V. Manikandan, P. Manikandan.
Approval of the final text: Velantina V, V. Manikandan, P. Manikandan.

VII. ACKNOWLEDGMENTS

The authors would like to extend their heartfelt thanks to the Department of Computer science and Engineering, Jain University, Bengaluru, for their support of this work.

VIII. REFERENCES

[1] D. Buffelli and F. Vandin, "Attention-based deep learning framework for human activity recognition with user adaptation," IEEE Sensors Journal, vol. 21, no. 12, pp. 13474–13483, 2022.

[2] T. Ahmad, L. Jin, X. Zhang, S. Lai, G. Tang, and L. Lin, "Graph convolutional neural network for human action recognition: A comprehensive survey," IEEE Transactions on Artificial Intelligence, vol. 2, no. 2, pp. 128–145, 2021.

[3] M. Duhme, R. Memmesheimer, and D. Paulus, “Fusion-GCN: Multimodal action recognition using graph convolutional networks,” in Proceedings of the DAGM German Conference on Pattern Recognition (GCPR), LNCS, vol. 13024, pp. 265–281, 2021.

[4] M. Korban, P. Youngs, and S. T. Acton, “A multi-modal transformer network for action detection,” Pattern Recognition, vol. 142, p. 109713, 2023.

[5] A. Omolaja, A. Otebolaku, and A. Alfoudi, “Context-aware complex human activity recognition using hybrid deep learning models,” Applied Sciences, vol. 12, no. 18, p. 9305, 2022.

[6] J. Y. Kwon and D. Y. Ju, “Living Lab-based service interaction design for a companion robot for seniors in South Korea,” Biomimetics, vol. 8, no. 8, p. 609, 2023.

[7] M. Liu, F. Meng, and Y. Liang, “Generalized pose decoupled network for unsupervised 3D skeleton sequence-based action representation learning,” Cyborg and Bionic Systems, Art. no. 0002, 2022.

[8] V. Mazzia, S. Angarano, F. Salvetti, F. Angelini, and M. Chiaberge, “Action transformer: A self-attention model for short-time pose-based human action recognition,” Pattern Recognition, vol. 124, p. 108487, 2021.

[9] H. P. Nguyen and B. Ribeiro, “Video action recognition collaborative learning with dynamics via PSO-ConvNet transformer,” Scientific Reports, vol. 13, no. 1, Art. no. 39744, 2023.

[10] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in Advances in Neural Information Processing Systems, vol. 27, pp. 568–576, 2014.

[11] J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new model and the kinetics dataset,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6299–6308, 2017.

[12] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, “Video action transformer network,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 244–253, 2019.

[13] K. Alomar, H. I. Aysel, and X. Cai, “CNNs, RNNs and transformers in human action recognition: A survey and a hybrid model,” Artificial Intelligence Review, vol. 58, Art. no. 387, 2025.

[14] S. Subramanian et al., “Attention-based deep learning for human activity recognition: A review,” Cognitive Computation, in press, 2025.

[15] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, “A closer look at spatiotemporal convolutions for action recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6450–6459, 2018.

[16] C. Feichtenhofer, H. Fan, J. Malik, and K. He, “SlowFast networks for video recognition,” International Journal of Computer Vision, vol. 130, pp. 290–312, 2022.

[17] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 677–691, 2021.

[18] Y. Zhang, Z. Xu, and L. Wang, “CNN–LSTM based video action recognition,” Pattern Recognition, vol. 124, p. 108414, 2022.
[19] S. Bai, J. Z. Kolter, and V. Koltun, “An empirical evaluation of temporal convolutional networks for sequence modeling,” Neurocomputing, vol. 452, pp. 191–206, 2022.

[20] W. Li, L. Wang, and H. Zhang, “Temporal convolutional modeling for video action recognition,” Neural Networks, vol. 156, pp. 205–217, 2023.

[21] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, “ViViT: A video vision transformer,” International Journal of Computer Vision, 2022.

[22] H. Han, H. Zeng, L. Kuang, and M. Xue, “Vision transformer-based human activity recognition,” Scientific Reports, vol. 14, Art. no. 65850, 2024.

[23] M. Liu, J. Wang, B. He, and L. Qu, “Hybrid CNN-transformer networks for human activity recognition,” Applied Sciences, vol. 13, no. 5, p. 2695, 2023.

[24] R. R. Dokkar, F. Chaieb, H. Drira, and A. Aberkane, “ConViViT: Convolution–vision transformer for human activity recognition,” Pattern Recognition Letters, vol. 171, pp. 55–63, 2023.

[25] S. Yan, Y. Xiong, and D. Lin, “Spatial-temporal graph convolutional networks for action recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 7, pp. 2435–2448, 2021.

[26] C. Zhang, Y. Wang, and D. Tao, “Graph-based temporal reasoning for video action recognition,” IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 8, pp. 4221–4234, 2023.

[27] J. Wang et al., “Multimodal video fusion for activity recognition,” Information Fusion, vol. 78, p. 102464, 2023.

[28] M. A. Khan et al., “Deep multimodal learning for video-based human activity recognition,” Expert Systems with Applications, vol. 226, p. 120059, 2024.

[29] R. Zhao, L. Chen, and S. Liu, “Benchmarking deep learning models for video human activity recognition,” Knowledge-Based Systems, vol. 295, p. 107826, 2024.

[30] Y. Chen, J. Li, and X. Yang, “Evolutionary optimized deep networks for action recognition,” Neurocomputing, vol. 513, pp. 153–166, 2023.

[31] P. Singh, R. Gupta, and G. Varshney, “Meta-heuristic optimization for deep human activity recognition models,” Applied Soft Computing, vol. 135, p. 110135, 2024.

[32] X. Li, B. Zhou, and Y. Sun, “Scalable transformer frameworks for large-scale video action recognition,” IEEE Transactions on Multimedia, vol. 27, pp. 456–469, 2025.