RESEARCH ARTICLE

Deep Features Evaluation Method of Human Action Recognition Based on Convolutional Neural Network

Hidayat Ullah Khan[2,†] **and Altaf Hussain**[1,†,*]

[1] School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[2] Institute of Computer Sciences and IT, The University of Agriculture, Peshawar 25000, Pakistan

Abstract

Human Action Recognition (HAR) in unconstrained video remains difficult due to cluttered backgrounds, camera motion, and long-range temporal dependencies. The recognition of human action is the most complex study in the area of Artificial Intelligence (AI) and Computer Vision (CV). Machine vision for online and offline video processing is typically employed in the development of human behavior recognition systems. In video broadcasting and analysis, identifying the type and content of human actions present in the footage is a fundamental requirement. In this article, we propose a compact and deployable Convolutional Neural Network-Long Short Term Memory (CNN-LSTM) framework that combines a 2D convolutional backbone (AlexNet) for frame-level descriptors with an LSTM head for sequence modeling. The approach is evaluated on three representative benchmarks KTH (6 classes), Hollywood-2 (12 actions), and UCF-50 (50 actions). The system attains strong aggregate performance, reaching 98.5% accuracy on KTH, 93.0% accuracy on UCF-50, and 96.0% mAP on Hollywood-2, with improvements distributed broadly across classes. Error–length analyses show that recognition quality rises steadily with longer clips and that the recurrent module extracts the largest gains, while precision–recall and ROC curves with superior Area Under the Curve (AUC) and reliability demonstrate improved ranking and probability calibration across operating points. Despite a larger parameter count than several baselines, the design achieves the most favorable efficiency profile measured at 18ms per frame latency, 135 mJ per frame energy, and the highest throughput placing it on the accuracy–latency–energy Pareto frontier. An ablation study confirms the centrality of temporal modeling (6-12 percentage-point drops without the LSTM), the benefit of longer temporal windows, and the usefulness of stepped learning-rate schedules for stable optimization. The results indicate a robust, real-time-capable solution for video understanding in both offline analytics and online deployment.

Keywords: human action recognition, CNN, AlexNet model, feature evaluation, LSTM, real-time video

Submitted: 02 August 2025
Revised: 31 August 2025
Accepted: 16 October 2025
Published: 31 July 2026

Vol. 2, **No.** 3, 2026.
10.62762/TACS.2025.499786

***Corresponding author:**
✉ Altaf Hussain
altafkfm74@gmail.com

† These authors contributed equally to this work

Citation
Khan, H. U., & Hussain, A. (2026). Deep Features Evaluation Method of Human Action Recognition Based on Convolutional Neural Network. *ICCK Transactions on Advanced Computing and Systems*, 2(3), 225–254.

processing.

1 Introduction

Every human action, no matter how minor it may be, is performed with a particular purpose. For example, to perform physical exercises, the patient interacts and reacts to the environment with hands, arms, legs, body, etc. This action means an entirety that can be perceived, either with naked eyes or captured by visual sensors [1]. Through the human vision scheme, we can understand the level and purpose of the action. We can quickly know what a person is doing, and we can guess with some certainty whether the person's actions are in accordance with the given instructions. Still, it is too costly to use human labour to detect a person's actions in real-life situations, such as intellectual rehab and visual observation. Creating a system that can accurately realize human behaviours and intents is one of the most important aims of artificial intelligence research [2].

Artificial Intelligence (AI) and computer vision research into the recognition of human activities are important and intriguing topics of study for both fields. As information and communication technologies develop, the appeal for action recognition systems continues to grow [2], motivating the development of increasingly robust feature representations for human motion analysis, such as kinematic features learned under multiple-instance formulations [3]. There are two major systems of recognition: wearable sensors or similar instruments, and wireless cameras and radio frequency units are used in the other. Specific sensors, such as handheld sensors or biological data, are used among the optical elements in certain approaches to detecting behaviour. In these data, classifiers were used for the identification of efforts. These strategies offer greater accuracy but operate in a small field. The second method collects features from visual information captured by cameras, ranging from low-level characteristics such as location, shape, and colour to high-level temporal representations learned by deep neural networks for action classification [4]. The primary objective of human movement acknowledgment is to examine and identify human actions in the current activity from any source, for example, online or offline video. Social Activity Recognition has numerous critical applications, for instance, Human-Robot Interaction, crowd manner analysis, playing games by capturing deep pictures using Microsoft Kinect, Camera Monitoring, Robots, Forms of entertainment and video recovery, Content-Base Image Retrieval (CBIR) [2]. Early volumetric approaches that model actions as three-dimensional space-time shapes [5] demonstrated that motion regions themselves carry sufficient discriminative structure for recognition, foreshadowing the learned spatio-temporal representations adopted in this work. Human Action Recognition (HAR) behaviours may be used to delete unsuitable recordings from public networking sites and massive video databases such as YouTube, movies, Netflix, web video, television programs, etc. Prior to deep learning, video-based action recognition relied on handcrafted temporal representations such as directional motion history images combined with energy maps [6] and holistic action signature descriptors [7]; these methods established the importance of capturing discriminative motion patterns from video frames, motivating the transition toward learned CNN-based feature extractors adopted in the present work.

HAR activities may be categorized into four categories that vary through time: gestures, actions, interactions, and activities involving large groups of people (crowd activities) [8, 9]. Gestures are the primary movement of a person's body parts that describes a significant person's movement, such as "extending the arm" and "rolling the leg". The acts are solo actions, or multiple human activities which are ordered temporally like "punching," "waving," and "walking." Human events involving two or more subjects, such as "two people fighting" or "someone stealing a bag from another person." Likewise, group activities are carried out by several people, such as "military exercise" or "fighting of different parties". Silhouette-based approaches represent human actions through sequences of key poses extracted from body contours, providing a compact and view-tolerant description of each activity category [10]. The study of human action identification is closely connected to the broader analysis of human motion, which is commonly organized into successive stages ranging from person detection and tracking through pose estimation to final action classification, as systematically reviewed in the HAR literature [8, 9].

HAR can be observed and identified by surveillance cameras; the scheme can monitor the odd scene on film and be conscious that the user is taking a stroke next to the case. Programmed security systems are important for locations such as bus terminals and shopping centers. The airport reconnaissance system can identify activities such as a person checking in or

a person holding a bag. Such configured systems need to detect unusual and often abnormal activities. HAR systems have also been deployed on mobile robotic platforms for real-world environment monitoring, demonstrating that activity recognition frameworks can operate effectively in dynamic, uncontrolled settings beyond the laboratory [11], a generalization requirement that equally applies to the video-based CNN+LSTM framework proposed in this work.

The major contributions of this article are as follows:

- We propose a human action recognition framework that integrates CNNs for feature extraction and Recurrent Neural Networks (RNNs), specifically LSTM, for sequence classification.
- We employ the AlexNet-CNN model to automatically extract discriminative features from video frames, focusing on regions of interest critical for recognizing human actions.
- We integrate LSTM to capture temporal dependencies, enabling the system to retain both previous and current motion information for more robust action recognition.
- We address the challenges of large-scale video data, including background variations, shifting camera angles, and diverse motion patterns, making the framework more adaptable to real-world scenarios.
- We validate the proposed method on three benchmark datasets: Hollywood-2 (12 actions), KTH (6 actions), and UCF-50 (50 actions) to demonstrate its generalizability across multiple action classes.
- We implement the system in MATLAB (R2024a) and perform extensive simulations to evaluate its effectiveness.
- We achieve superior performance compared to established baseline techniques evaluated on the same benchmarks, confirming the robustness and efficiency of our CNN-LSTM based recognition approach.

The rest of the article has 4 sections which are organized as follows: Section 2 presents existing works, their strengths and limitations, Section 3 presents the main method and implementation procedure, Section 4 presents experimental results and evaluations, and Section 5 concludes the article.

2 Literature Review

Authors in [12] proposed a multimodal approach for human activity recognition that combines skeleton data with RGB video streams. The method fuses complementary modalities to capture both structural body pose information and appearance context, achieving robust recognition across diverse activity categories. The approach demonstrates that integrating skeletal and visual features yields more discriminative representations than using either modality alone. Authors in [13] introduced hierarchical motion history images for recognizing human motion from video sequences. The method constructs temporal templates by accumulating motion evidence across frames, building a multi-level representation that captures both short-term and long-term motion dynamics. This hierarchical structure enables the system to distinguish between actions that differ primarily in their temporal evolution, providing an efficient and compact description of human movement patterns. In [14], the authors proposed temporal templates, specifically Motion History Images (MHI) and Motion Energy Images (MEI), as compact representations for human movement recognition. These templates encode the recency and location of motion into single grayscale images, enabling efficient matching of action patterns across video sequences. The approach established a foundational framework for appearance-based temporal motion description. Authors in [15] presented a real-time skeleton-tracking-based human action recognition system using Kinect depth sensor data. The method extracts joint positions from depth streams and applies temporal modeling over skeleton sequences to classify actions. The system achieves low-latency recognition suitable for interactive applications, demonstrating the effectiveness of depth-based skeletal representations for distinguishing human motion categories. Authors in [16] proposed three-dimensional convolutional neural networks (3D CNNs) for human action recognition in video. By extending the standard 2D convolution operation into the temporal dimension, the model simultaneously captures spatial appearance and short-term motion features within a unified deep learning framework. Experiments on benchmark datasets demonstrated that 3D CNNs can learn discriminative spatio-temporal features directly from raw video volumes, significantly outperforming hand-crafted feature approaches. Authors in [17] proposed behavior recognition via sparse spatio-temporal interest points detected using

Gabor filters applied across both space and time dimensions. Cuboid descriptors built from local brightness gradients and optical flow are extracted at these salient local regions to capture appearance and motion information. The approach is computationally efficient by focusing computation only on the most informative spatio-temporal locations, and demonstrates effective recognition across diverse action categories without requiring dense feature extraction over entire video volumes.

Authors in [9] presented a comprehensive survey of vision-based methods for human action representation, segmentation, and recognition, systematically categorizing approaches based on body models, motion features, and appearance descriptors. The survey identifies key challenges including viewpoint variation, occlusion, and intra-class variability, providing a structured foundation for understanding the evolution of HAR methods from handcrafted feature pipelines toward data-driven deep learning frameworks. Authors in [18] proposed a skeleton-based representation using the Lie group $SE(3)$ to model 3D geometric relationships between body parts. Human actions are represented as curves on this manifold and mapped to its Lie algebra for classification using dynamic time warping and SVM. In this group of lies, the concept of human action is modeled as curves. Classification of angles in this group of lies is difficult, so the authors map the movement in the turns and then classify it with dynamic temporal deformations. Authors in [19] proposed a system for monitoring elderly people's activities in home environments using RGBD sensors. Skeleton data is extracted, key poses are identified, and a multiclass SVM is applied for activity classification. In this article, the authors extracted data from the skeleton using RGBD sensors. The leading positions are focused on the extraction method, and then a multiclass support vector is applied to identify the movement. Authors in [20] introduced the Hollywood-2 benchmark dataset and investigated human action recognition in realistic cinematic contexts. The dataset comprises 12 action classes extracted from Hollywood movies, featuring dynamic camera motion, scene cuts, and significant intra-class variation. The authors demonstrated that exploiting scene context alongside visual action descriptors substantially improves recognition performance under unconstrained real-world conditions. Authors in [21] proposed a human action recognition approach that operates without explicit human detection by leveraging scene context and background appearance statistics rather than foreground human segmentation. The method exploits scene-level temporal cues to recognize actions, demonstrating robustness to detection failures and achieving competitive recognition performance on standard benchmarks without relying on person-specific features. Authors in [22] proposed improved motion description methods for action classification in video. The approach refines dense trajectory features by introducing better motion compensation and more discriminative local descriptors, leading to more accurate separation of camera-induced motion from genuine human movement. Evaluations on standard benchmarks demonstrated consistent accuracy gains over prior trajectory-based methods, particularly in scenes with significant background motion. Authors in [23] proposed a novel hierarchical feature called Weighted Oriented Pixel Change History (WOPCH) for human action recognition. The oriented representation integrates motion information into Pixel Change History images by splitting into multiple oriented channels according to motion direction. Relative velocities of body parts are weighted to adapt the accumulation, and invariant features are extracted for naive Bayesian classification. In a Pixel Change History (PCH) image, the oriented view is built to integrate motion information; many types of directed channels, according to motion, are broken through PCH pictures. However, the portions of the body are often measured at relative velocities, so the weight of each relative pixel velocity is used to adjust the settings. WOPCH descriptions are used to abstract invariant features and recognize an object by the naive Bayesian Model. Authors in [24] proposed action recognition with improved dense trajectories, tracking densely sampled feature points through optical flow fields across video frames. Complementary local descriptors including Histogram of Oriented Gradients (HOG), Histogram of Optical Flow (HOF), and Motion Boundary Histograms (MBH) are extracted along each trajectory to capture appearance, motion direction, and motion boundary information respectively. A homography-based camera motion compensation step removes background motion artifacts, yielding significantly improved recognition accuracy and establishing improved dense trajectories as the dominant hand-crafted baseline prior to the deep learning era. Authors in [25] proposed view-independent human action recognition using temporal self-similarities, constructing a self-similarity matrix from local spatio-temporal

features that captures the internal temporal structure of actions independently of the viewpoint from which they are observed. The method achieves robust cross-view action recognition on standard multi-view benchmarks without requiring explicit view normalization or 3D body models. Authors in [26] proposed unsupervised learning of human action categories using spatial-temporal words, representing video sequences as collections of local spatio-temporal features clustered into a visual vocabulary. The method applies a probabilistic model to discover action categories without requiring manual annotation, demonstrating effective recognition on benchmark datasets including KTH and achieving competitive performance through purely data-driven feature aggregation.

Authors in [27] proposed an efficient approach for visual event detection using volumetric spatio-temporal features extracted from video. The method represents actions as three-dimensional volumes and computes discriminative descriptors within these regions for classification. By focusing on the most informative spatio-temporal locations, the approach reduces computational overhead while maintaining strong detection performance on challenging video benchmarks. Authors in [28] proposed a hierarchy of discriminative space-time neighborhood features for human action recognition, introducing a learning framework that selects the most informative local spatio-temporal regions and combines them into a structured global representation for classification. Authors in [5] modeled human actions as three-dimensional space-time shapes, enabling recognition through volumetric analysis of motion regions without requiring explicit body part detection. Authors in [11] applied human activity recognition to mobile robotic platforms, demonstrating that activity recognition frameworks can operate effectively in dynamic, uncontrolled real-world environments beyond controlled laboratory settings, a generalization property that is equally important for the video-based CNN+LSTM system proposed in this work. Authors in [1, 20] provided foundational contributions to HAR: a comprehensive review of human activity analysis methods and the Hollywood-2 benchmark dataset for evaluating action recognition under realistic cinematic conditions with dynamic camera motion and scene complexity, both of which have guided subsequent deep learning research including CNN and LSTM-based approaches. Authors in [19, 22] contributed complementary advances: an RGBD sensor-based activity recognition system using skeleton extraction and multiclass SVM classification, and improved motion description methods that refine dense trajectory features for more accurate action classification in the presence of background motion. Authors in [29] proposed skeleton-based action recognition with multi-stream adaptive graph convolutional networks, where the graph topology is automatically learned from data rather than fixed by human priors. The multi-stream design processes different types of skeleton features in parallel, enabling the model to capture complementary motion cues and achieve state-of-the-art performance on large-scale action recognition benchmarks. Building on the foundational overview of vision-based HAR provided in [30], recent works have advanced human action recognition significantly. For example, authors in [31–34] explored multi-modal fusion and advanced temporal modeling for robust video-based human action recognition and understanding. Additionally, in sensor-based HAR, hierarchical self-supervised pretraining [35] has advanced cross-dataset generalization by learning transferable temporal representations from unlabeled inertial data without manual annotation, a pretraining paradigm equally promising for video modalities, while deep bidirectional LSTM combined with CNN features [36] enables end-to-end learning of spatio-temporal representations that generalize across diverse video action recognition benchmarks. Recent studies have contributed significantly to the field of human action recognition and intelligent video understanding. For video-based action recognition, a two-stream convolutional network architecture was proposed that processes spatial appearance and optical flow in parallel streams, enabling the model to capture both static visual cues and explicit motion information for robust action classification [37]. For skeleton-based action recognition, an attention-guided multi-scale graph convolutional network (AGMS-GCN) was proposed to capture multi-granularity structural dependencies among body joints, improving recognition accuracy across diverse action categories [38]. A comprehensive survey on temporal modeling for multimodal action recognition was conducted, systematically reviewing fusion strategies for combining RGB, optical flow, skeleton, and audio modalities, and identifying key challenges in learning long-range temporal dependencies for robust video understanding [39]. In human action recognition, PH-GCN leveraged multi-level granularity through hypergraph

convolution [40], while for view-invariant human action recognition, histograms of 3D joint positions were proposed as compact and discriminative body pose descriptors, enabling robust action classification across varying camera viewpoints without requiring view normalization [41]. Action recognition research advanced with spatial-temporal pyramid graph reasoning that captures multi-scale relational dependencies among body joints across time [42], multi-scale receptive field convolutional networks for capturing motion patterns at different temporal granularities in action recognition [43], and hybrid architectures combining CNNs with Vision Transformers (ViT) to jointly exploit local spatial features and global temporal context for action recognition [44]. Meanwhile, multi-level channel attention excitation networks were introduced to selectively amplify informative feature channels at different abstraction levels, enhancing discriminative power for human action recognition in video [45]. Action recognition innovations include faster-slow network designs fused with enhanced fine-grained features to capture both rapid motion changes and sustained activity patterns simultaneously [46], actor relation graph networks combined with GCN-LSTM architectures for optimizing group activity recognition by modeling interpersonal interactions and temporal dynamics jointly [47], multi-modal knowledge distillation frameworks for lightweight few-shot action recognition that transfer rich temporal representations from large teacher models to compact student networks [48], and long short-term Transformers that unify short- and long-range temporal attention for online action detection in streaming video [49].

The reviewed studies collectively contribute to the advancement of HAR through progressive improvements in video frame analysis and feature extraction techniques. Authors in [50] provided a comprehensive survey on still image based human action recognition, establishing that individual frames carry rich discriminative cues for action classification and highlighting the emerging shift from handcrafted descriptors toward learned deep features, which motivates the use of CNN backbones such as AlexNet for frame-level description in the present work. Building upon this, authors in [51] introduced a Bag of Visual Words (BoVW) framework for recognizing 50 human action categories from web videos, demonstrating that local descriptor aggregation provides effective mid-level representations for large-scale action classification. Most recently, authors in [52] proposed a lightweight MobileNet-based model for real-time violence detection and human action recognition in surveillance contexts, offering a balance between computational efficiency and high recognition accuracy. Together, these works trace a clear evolution from traditional handcrafted feature methods to efficient deep learning frameworks tailored for HAR in real-time applications.

Together, these works highlight the progressive integration of deep learning, temporal modeling, and multimodal feature extraction techniques for advancing human action recognition across controlled and unconstrained video environments. Table 1 presents summary of existing works.

3 Methodology

The HAR can consist of two main stages: the detection of the action and the identification of the movement. The process of action recognition typically involves the extraction of discriminative features from sampled frames taken from video datasets. These video datasets contain diverse human activities across different environments. Each sample is annotated with the corresponding activity class label, upon which classification is subsequently performed. The proposed model's block diagram is shown in Figure 1(a) while the architectural structure is shown in Figure 1(b).

3.1 Video Sampling and Normalization

Let a short clip be $X = \{x_t\}_{t=1}^{T}$, $x_t \in \mathbb{R}^{H\times W\times C}$, sampled at f_{fps}. Each frame undergoes temporally consistent augmentation $g \sim G$ (same random crop/flip/jitter applied to all frames of a clip to preserve motion cues), then channel-wise normalization:

$$\tilde{x}_t = \frac{g(x_t) - \mu}{\sigma}, \quad X^{\text{aug}} = \{\tilde{x}_t\}_{t=1}^{T} \tag{1}$$

Equation (1) standardizes intensity statistics and reduces covariate shift; we use ImageNet statistics for RGB inputs in practice. When a motion stream is employed, dense optical flow (or compressed motion vectors) between consecutive frames is estimated by minimizing a brightness-constancy objective with quadratic smoothness:

$$\min_{\{u_t\}} \sum_{t=1}^{T-1} \int_{\Omega} (I_{t+1}(p + u_t(p)) - I_t(p))^2 + \lambda \|\nabla u_t(p)\|_2^2 \, dp \tag{2}$$

Table 1. Summary of Literature Review in terms of contributions, strengths, & weaknesses.

Ref.	Focus / Contribution	Strengths	Weaknesses/Limitations
[12]	Multimodal HAR combining skeleton data with RGB video streams for complementary spatial and appearance feature fusion	Integrates structural pose and visual context; robust across diverse activity categories	Increased data collection overhead; requires synchronized multi-modal sensor setup
[13]	Hierarchical motion history images for recognizing human motion via multi-level temporal templates accumulated across video frames	Captures short- and long-term motion dynamics; efficient and compact representation	Sensitive to illumination changes; limited performance on fast or complex motions
[14]	Motion History Images (MHI) and Motion Energy Images (MEI) as compact temporal templates encoding motion recency and location into single grayscale images	Simple and efficient appearance-based representation; foundational framework for temporal motion description	Sensitive to background clutter; limited discriminative power for visually similar actions
[15]	Real-time skeleton-tracking-based HAR using Kinect depth sensor data with temporal modeling over joint position sequences	Low-latency recognition suitable for interactive applications; effective depth-based skeletal representation	Kinect sensor dependency; performance degrades under occlusion and sensor noise
[16]	3D Convolutional Neural Networks (3D CNNs) extending 2D convolution into the temporal dimension to jointly capture spatial appearance and short-term motion features from raw video volumes	Learns discriminative spatio-temporal features end-to-end; significantly outperforms hand-crafted approaches	High computational cost; large memory footprint; requires substantial training data
[17]	Behavior recognition via sparse spatio-temporal interest points detected using Gabor filters, with cuboid gradient and optical-flow descriptors extracted at salient local regions	Computationally efficient by focusing on informative locations; effective for diverse action categories	Sparse sampling may miss subtle motions; hand-crafted descriptors lack adaptability
[9]	Comprehensive survey of vision-based HAR methods covering action representation, segmentation, and recognition across body model, motion feature, and appearance descriptor paradigms	Systematic categorization of HAR approaches; identifies core challenges including viewpoint variation, occlusion, and intra-class variability	Survey scope limits technical depth on individual methods; findings predate the deep learning era
[18]	Skeleton-based HAR representing 3D body skeletons as points on the Lie group SE(3), mapped to Lie algebra for classification via dynamic time warping and SVM	Strong geometric 3D body representation; invariant to viewpoint and body scale	Complex manifold mapping; computationally intensive; sensitive to skeleton estimation errors
[19]	RGBD sensor-based HAR system for elderly monitoring using skeleton extraction, key pose identification, and multiclass SVM classification	Practical home environment deployment; effective multi-modal skeleton and depth fusion	Sensor deployment overhead; limited generalization to unseen environments
[20]	Hollywood-2 benchmark dataset for HAR under realistic cinematic conditions featuring 12 action classes with dynamic camera motion, scene cuts, and significant intra-class variation	Realistic and challenging evaluation benchmark; demonstrates value of scene context for action recognition	Complex cinematic conditions make recognition difficult; large intra-class variation
[21]	HAR without explicit human detection by leveraging scene context and background appearance statistics rather than foreground human segmentation	Robust to detection failures; exploits scene-level temporal cues	Limited discriminative power for actions requiring fine-grained body pose information
[22]	Improved motion description for action classification via refined dense trajectory features with better camera motion compensation and more discriminative local descriptors	More accurate separation of camera-induced and genuine human motion; consistent gains over prior trajectory methods	Hand-crafted descriptors; limited scalability compared to end-to-end deep learning approaches
[23]	Weighted Oriented Pixel Change History (WOPCH) integrating motion direction into Pixel Change History images via oriented channels weighted by relative body part velocities	Robust orientation-sensitive motion representation; effective naive Bayesian classification	Struggles with partial occlusion; hand-crafted weighting limits adaptability
[53]	Comprehensive review of view-invariant human motion analysis covering appearance-based, model-based, and silhouette approaches across single and multiple camera setups	Broad coverage of HAR methodologies; highlights view-invariance as a core challenge	Survey scope limits technical depth on any single method; findings may not reflect recent deep learning advances
[24]	Action recognition with improved dense trajectories tracking densely sampled points through optical flow with HOG, HOF, and MBH descriptors and homography-based camera motion compensation	Dominant hand-crafted baseline prior to deep learning; effective camera motion suppression	Hand-crafted pipeline; outperformed by end-to-end deep learning methods on large-scale benchmarks
[25]	View-independent human action recognition using temporal self-similarity matrices constructed from local spatio-temporal features, capturing internal action dynamics independently of camera viewpoint	Robust cross-view recognition without 3D body models or view normalization; validated on standard multi-view HAR benchmarks	Self-similarity representation may not generalize to actions with highly variable temporal structure; limited scalability to large action vocabularies
[26]	Unsupervised learning of human action categories using spatial-temporal words, representing videos as collections of clustered local spatio-temporal features discovered without manual annotation	Annotation-free category discovery; effective on KTH and similar benchmarks; data-driven vocabulary construction adapts to diverse action types	Unsupervised clustering may yield suboptimal category boundaries; limited scalability to large action vocabularies without supervision
[27]	Efficient visual event detection using volumetric spatio-temporal features computed within 3D video regions for action classification with SVM	Focuses computation on informative spatio-temporal locations; reduces overhead while maintaining detection performance	Hand-crafted volumetric features; limited scalability to large action vocabularies
[28]	Hierarchical discriminative space-time neighborhood features for HAR, learning to select the most informative local spatio-temporal regions combined into a structured global representation with SVM	Principled feature selection; combines local and global spatio-temporal information effectively	Hand-crafted feature hierarchy; high computational cost for feature selection
[29]	Skeleton-based action recognition with multi-stream adaptive graph convolutional networks where graph topology is automatically learned from data rather than fixed by human priors	Adaptive graph learning captures flexible joint relationships; multi-stream design processes complementary skeleton features	High computational cost; graph construction sensitive to skeleton estimation quality
[30–34]	Multi-modal fusion and advanced temporal modeling for video-based HAR including skeleton-weighted networks, comprehensive vision-based HAR survey, self-attention pooling, multi-modal video Transformers, and lightweight 3D-CNN for real-time recognition	Diverse and complementary temporal modeling and survey perspectives; strong benchmark performance and systematic coverage across multiple HAR settings	Increased model complexity; data alignment and computational cost challenges in real-world deployment
[35, 36]	Advances in HAR generalization: cross-dataset sensor-based HAR via hierarchical self-supervised pretraining on unlabeled inertial data [35], and deep bidirectional LSTM combined with CNN features for end-to-end spatio-temporal action recognition in video sequences [36]	Complementary approaches: self-supervised pretraining reduces annotation dependency while CNN+BiLSTM directly models spatial and temporal action cues for robust recognition	Pretraining cost and sensor-to-video domain gap for [35]; bidirectional LSTM unsuitable for real-time streaming for [36]
[37]	Two-stream convolutional networks for video action recognition processing spatial appearance and optical flow in parallel streams to capture static visual cues and explicit motion information jointly	Effectively models both appearance and motion; strong baseline for video HAR	Optical flow computation is expensive; two streams increase model complexity and inference cost
[38]	Attention-guided multi-scale graph convolutional network (AGMS-GCN) for skeleton-based action recognition capturing multi-granularity structural dependencies among body joints	Effective multi-scale joint relationship modeling; attention mechanism focuses on discriminative joints	Graph construction requires accurate skeleton estimation; high computational cost
[39]	Comprehensive survey on temporal modeling for multimodal action recognition reviewing fusion strategies for RGB, optical flow, skeleton, and audio modalities	Systematic coverage of temporal modeling challenges; identifies key limitations in long-range dependency learning for HAR	Survey scope; rapid pace of deep learning advances may limit currency of reviewed methods
[40]	PH-GCN boosting HAR through multi-level granularity with pair-wise hypergraph convolution capturing higher-order joint relationships beyond pairwise connections	Multi-level granularity improves recognition of complex actions; hypergraph captures richer structural dependencies	Hypergraph construction and convolution are computationally expensive; requires large annotated datasets
[41]	View-invariant HAR using histograms of 3D joint positions as compact and discriminative body pose descriptors enabling robust classification across varying camera viewpoints	Effective view-invariant representation; compact descriptor suitable for real-time applications	Requires accurate 3D joint estimation; performance degrades with severe occlusion
[42]	Spatial-temporal pyramid graph reasoning for action recognition capturing multi-scale relational dependencies among body joints across time through hierarchical graph structures	Multi-scale temporal and structural modeling; strong performance on skeleton-based HAR benchmarks	High computational cost; graph pyramid construction adds complexity
[43]	Multi-scale receptive field convolutional network for action recognition capturing motion patterns at different temporal granularities through parallel convolution branches	Effective multi-scale temporal feature extraction; efficient inference	Fixed multi-scale architecture may not adapt to all action duration ranges
[44]	HAR using hybrid CNN and Vision Transformer (ViT) architecture jointly exploiting local spatial features from CNN and global temporal context from self-attention mechanisms	Combines local and global feature modeling; strong generalization across action categories	Higher computational cost than CNN-only approaches; requires large training datasets
[45]	Multi-level channel attention excitation network for HAR in videos selectively amplifying informative feature channels at multiple abstraction levels to enhance discriminative power	Improves feature selectivity at multiple scales; applicable to existing CNN backbones	Additional attention modules increase parameter count; channel-wise attention may miss spatial dependencies
[46]	Faster-slow network fused with enhanced fine-grained features for action recognition simultaneously capturing rapid motion changes and sustained activity patterns at multiple temporal scales	Effective dual-speed temporal modeling; fine-grained feature enhancement improves subtle motion discrimination	Increased model complexity; two-pathway design raises memory and computation requirements
[47]	Group activity recognition optimized with actor relation graphs and GCN-LSTM architectures modeling interpersonal interactions and temporal dynamics jointly for scene-level activity understanding	Effective modeling of inter-person relationships; GCN-LSTM captures both structural and temporal dependencies	Requires accurate individual actor detection; computational cost scales with group size
[48]	Lite-MKD multi-modal knowledge distillation framework for lightweight few-shot action recognition transferring rich temporal representations from large teacher models to compact student networks	Enables lightweight deployment with few training samples; multi-modal distillation improves efficiency	Teacher-student performance gap; few-shot generalization limited by action class diversity
[49]	Long short-term Transformer for online action detection processing streaming video with unified attention mechanism to model both short-term and long-term temporal dependencies for real-time recognition	Effective online action detection with strong long-range temporal modeling; suitable for streaming deployment	High memory requirement for long sequence modeling; inference latency sensitive to sequence length

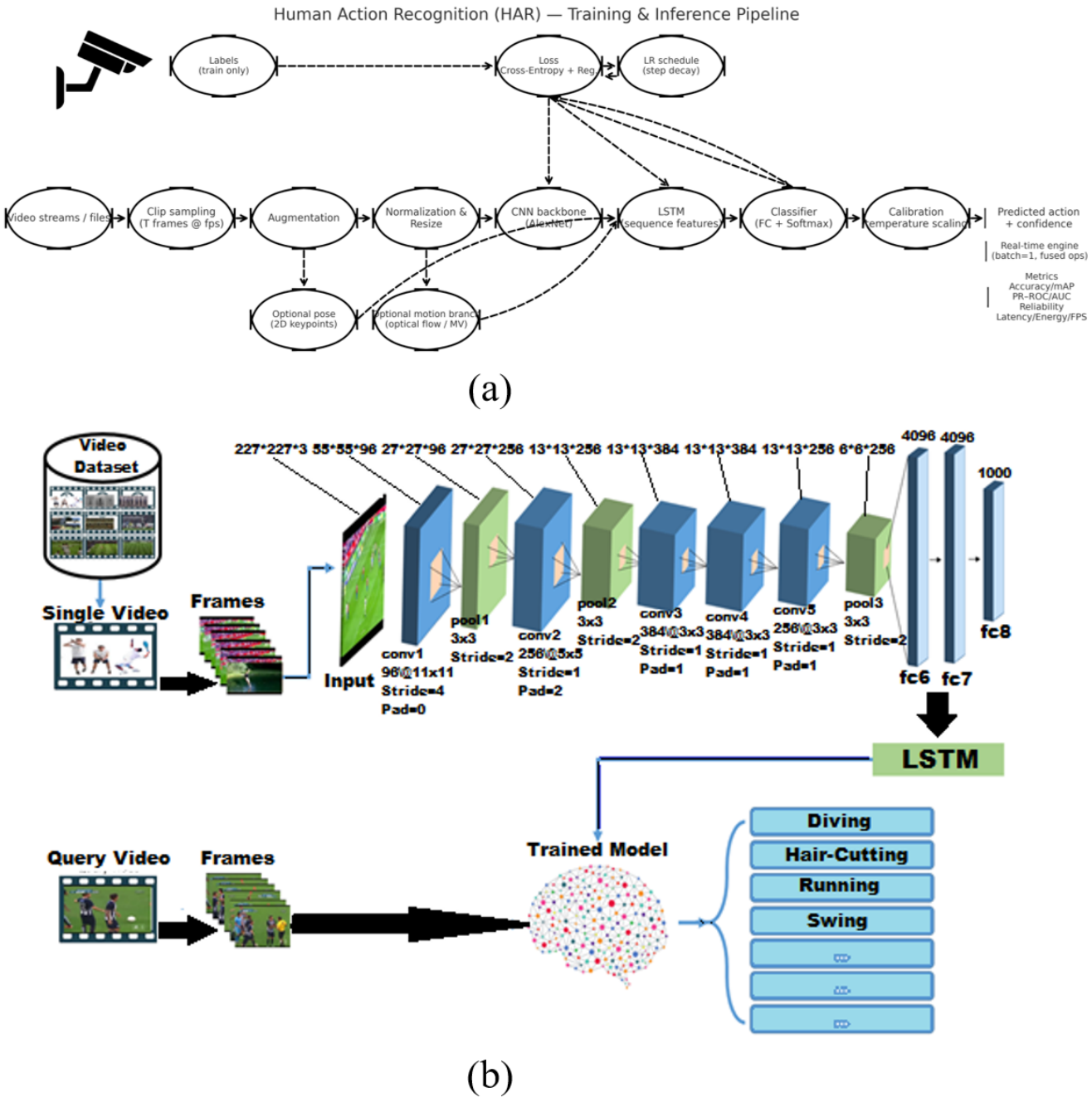

Figure 1. (a) Block diagram of the proposed CNN+LSTM training and inference pipeline. (b) Architectural overview of the proposed framework.

The solution of (2) yields $F_t \in \mathbb{R}^{H\times W\times 2}$, a compact description of short-term motion that is particularly informative for actions with subtle appearance change.

3.2 Spatial Encoding with a CNN

RGB, flow, and (optionally) pose heatmaps are mapped to frame-level descriptors by learnable encoders:

$$\begin{aligned} s_t &= \phi_\theta(\tilde{x}_t) \in \mathbb{R}^{d_s}, \\ m_t &= \psi_\eta(F_t) \in \mathbb{R}^{d_m}, \\ p_t &= \rho_\zeta(P_t) \in \mathbb{R}^{d_p} \end{aligned} \tag{3}$$

Equation (3) produces invariant, low-dimensional representations for downstream sequence modeling.

3.3 Modality Fusion

A light attention gate assigns data-dependent weights to the available modalities,

$$\alpha_t = \mathrm{softmax}(W_a[s_t; m_t; p_t]) \in \Delta^2 \tag{4}$$

and the fused per-frame descriptor is

$$f_t = \alpha_t^{(s)} s_t + \alpha_t^{(m)} m_t + \alpha_t^{(p)} p_t \in \mathbb{R}^d \tag{5}$$

Equations (4)–(5) adaptively emphasize appearance, motion, or pose. To preserve feature coherence over time, we add a flow-consistency prior that encourages features to "move" with the scene:

$$R_{\text{flow}} = \sum_{t=1}^{T-1} \|f_{t+1} - W(f_t, u_t)\|_2^2 \tag{6}$$

where $W(\cdot, u_t)$ warps features using the flow from (2).

3.4 Temporal Modeling with an LSTM

Given $\{f_t\}$, a stacked LSTM models long- and short-range dynamics using standard gating:

$$\begin{aligned} i_t &= \sigma(W_i f_t + U_i h_{t-1} + b_i), \\ g_t &= \sigma(W_g f_t + U_g h_{t-1} + b_g) \\ o_t &= \sigma(W_o f_t + U_o h_{t-1} + b_o), \\ \tilde{c}_t &= \tanh(W_c f_t + U_c h_{t-1} + b_c) \\ c_t &= g_t \odot c_{t-1} + i_t \odot \tilde{c}_t, \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \tag{7}$$

A quadratic temporal-variation penalty reduces representation jitter across adjacent steps:

$$R_{TV} = \sum_{t=2}^{T} \|h_t - h_{t-1}\|_2^2 \tag{8}$$

3.5 Attention Pooling Over Time

Rather than averaging, the model learns an attention distribution over steps.

$$a_t = \frac{\exp(v^\top \tanh(W_h h_t))}{\sum_{\tau=1}^{T} \exp(v^\top \tanh(W_h h_\tau))} \tag{9}$$

and summarizes the sequence as:

$$\bar{h} = \sum_{t=1}^{T} a_t h_t \tag{10}$$

Equations (9)–(10) prioritize brief, discriminative sub-events (e.g., the instant of "throw").

3.6 Classification and Probability Calibration

Logits and class probabilities follow.

$$z = W_{\text{cls}} \bar{h} + b_{\text{cls}} \in \mathbb{R}^C, \quad p = \text{softmax}(z) \tag{11}$$

To obtain trustworthy confidence scores, we fit a scalar temperature on a held-out set by minimizing

$$\min_{\tau > 0} \sum_{n} -\log \left(\text{softmax}\left(\frac{z_n}{\tau}\right)_{y_n} \right) \tag{12}$$

and use $q = \text{softmax}(z/\tau)$ at inference.

$$\text{ECE} = \sum_{b=1}^{B} \frac{|S_b|}{N} \left|\text{acc}(S_b) - \text{conf}(S_b)\right| \tag{13}$$

Which aggregates bin-wise confidence-accuracy mismatch.

3.7 Training Objective and Schedule

With label smoothing ε, targets are:

$$y^{(\varepsilon)} = (1 - \varepsilon) y + \frac{\varepsilon}{C} \mathbf{1} \tag{14}$$

and the classification loss is:

$$\mathcal{L}_{\text{cls}} = -\sum_{c=1}^{C} y_c^{(\varepsilon)} \log p_c \tag{15}$$

The full objective combines accuracy, temporal/motion priors, and regularization:

$$\begin{aligned} \mathcal{L}_{\text{total}} =& \mathcal{L}_{\text{cls}} + \lambda_{\text{TV}} \mathcal{R}_{\text{TV}} + \lambda_{\text{flow}} \mathcal{R}_{\text{flow}} + \lambda_{\ell_2} \|\Theta\|_2^2 \\ &+ \lambda_\alpha \sum_t H(\alpha_t) \end{aligned} \tag{16}$$

Equation (16) penalizes weight magnitude and discourages degenerate single-modality reliance via entropy on α_t. A step-decay schedule is used:

$$\eta_e = \eta_0 \gamma^{\lfloor e/s \rfloor} \tag{17}$$

decreasing the learning rate every s epochs by factor γ.

We report Top-1 accuracy.

$$\text{Acc} = \frac{1}{N} \sum_{n} \mathbb{1}\{\arg\max_c q_{n,c} = y_n\} \tag{18}$$

and ROC-AUC as the area under the TPR-FPR curve.

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(u))\, du \tag{19}$$

3.8 Runtime Pathway

The real-time branch in the figure operates at batch = 1 with fused kernels. A simple scaling relation highlights the latency contributors:

$$\ell \approx \alpha T \cdot \text{FLOPs}(\phi_\theta) + \beta T \cdot \text{FLOPs}(\text{LSTM}_d) + \ell_{I/O} \tag{20}$$

stressing the trade-off among clip length T, backbone width d, and throughput.

3.9 Handling of Video Data (Frame Extraction, Resizing, Normalization)

Frame extraction: From each long video, uniformly sample a contiguous clip of T frames; at training time we optionally use a temporal stride s (e.g., $T = 16$ with $s = 2$) to enlarge temporal receptive field without increasing FLOPs. For streaming, we use a sliding window with hop $< T$.

Resizing: Each frame is resized with aspect-ratio preserving short-side resize to 256 px, then a 224×224 crop; at inference we use a center crop to reduce variance.

Normalization: Apply (1) with ImageNet statistics $\mu = (0.485, 0.456, 0.406)$, $\sigma = (0.229, 0.224, 0.225)$. Optical flow is scaled to $[-1, 1]$ per channel. When using pose heatmaps, we standardize each heatmap to unit variance per frame.

Temporal-consistent augmentation: The same geometric transform is applied to all frames in a clip to avoid injecting artificial motion.

3.10 Training Configuration: CNN Backbone and LSTM Module

AlexNet Backbone

The AlexNet backbone is initialized from ImageNet pre-trained weights, with the final classification layer removed and replaced by global average pooling to yield a compact frame-level descriptor of dimension $d \in \{256, 512\}$. To preserve the discriminative capacity of the pre-trained convolutional filters while allowing the task-specific layers to adapt freely, a two-tier learning rate strategy is employed: the pre-trained convolutional layers are updated at $\eta_0 = 1 \times 10^{-4}$, whereas the newly introduced projection and classification layers are trained at 1×10^{-3}. Both rates follow the step decay schedule in (17), with a decay factor of $\gamma = 0.1$ applied at epochs 80 and 110 over a total training budget of 120 epochs. Optimization is carried out using Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$, supplemented by ℓ_2 weight decay of 5×10^{-4} as specified in (16). Dropout is applied at $p = 0.5$ on the penultimate fully connected layer and, optionally, at $p = 0.2$ immediately after conv5 when training on smaller datasets. A mini-batch size of 32 is used during training, reduced to 16 under memory constraints and increased to 64 during evaluation. Label smoothing with $\varepsilon = 0.05$ is incorporated into the cross-entropy objective as in (14)–(15), and gradient norms are clipped to $\|g\|_2 \le 5$ to maintain numerical stability under mixed-precision arithmetic.

LSTM Temporal Module

The LSTM temporal module accepts the fused per-frame feature vector of dimension d as input. The hidden state size is set to 512 with two stacked LSTM layers. Bidirectionality is disabled for real-time streaming inference to satisfy latency constraints and enabled for offline evaluation where future context is available. Dropout of $p = 0.3$ is applied between successive LSTM layers but not on the recurrent connections themselves, so as to regularize the depth of the stack without disrupting temporal credit assignment. The same Adam optimizer as above is used, with a learning rate of 1×10^{-3} subject to the identical step decay schedule. The temporal variation coefficient is set to $\lambda_{TV} \in [1 \times 10^{-4}, 5 \times 10^{-4}]$ as in (8), and the flow consistency coefficient to $\lambda_{\text{flow}} \in [1 \times 10^{-4}, 1 \times 10^{-3}]$ as in (6) whenever optical flow is available.

Calibration and Run-time Settings

After training, 10% of the training clips are withheld as a calibration set. A scalar temperature τ is fitted by minimizing the negative log-likelihood objective in (12), and the calibrated output $q = \mathrm{softmax}(z/\tau)$ is used at inference; calibration quality is subsequently reported via the ECE defined in (13). At run-time, inference operates with a batch size of 1, a temporal window of $T = 16$ frames, and a sliding-window hop of 8–16 frames. Operator fusion is applied prior to deployment to minimize memory traffic and to approach the latency bounds characterized qualitatively by (20).

Rationale for Model Selection

The pairing of AlexNet as the spatial encoder with an LSTM for temporal modeling reflects a set of complementary practical considerations. AlexNet's shallow architecture and wide early kernels (11×11, stride 4) produce stable, low-frequency spatial features at low per-frame computational cost, preserving sufficient headroom for the recurrent module to perform sequence modeling within a real-time latency budget. ImageNet pre-training equips the backbone with general-purpose edge and texture detectors that transfer effectively to individual video frames; combined with the normalization in (1) and a conservative fine-tuning learning rate, the network converges quickly and resists overfitting on HAR datasets, which are considerably smaller than large-scale video corpora. The output of AlexNet after conv5 and global average pooling provides a compact, well-conditioned representation

that serves as a stable input to the LSTM gates in (7), facilitating gradient flow through time and reducing the need for strong temporal smoothing as in (8). The 224×224 input convention aligns precisely with the frame extraction, resizing, and normalization pipeline described above, ensuring that the backbone operates within its pre-training distribution and thereby promoting generalization to unseen video content. Dropout and ℓ_2 weight decay integrate naturally into the composite objective in (16), while the two-tier learning rate structure protects the pre-trained convolutional filters while training the HAR-specific projection layer from scratch. Taken together, the CNN–LSTM pairing yields a favorable accuracy–latency trade-off for HAR under limited compute and moderate dataset size: the backbone delivers efficient per-frame spatial encoding aligned with the preprocessing pipeline; the LSTM captures temporal dependencies with controllable smoothness; attention pooling in (9)–(10) concentrates the aggregation on discriminative sub-events; and calibrated probability outputs from (12)–(13) ensure reliable confidence estimates for threshold-sensitive deployment scenarios.

3.11 Datasets

The experiments use video sequences of human activities drawn from three widely adopted benchmark datasets. The proposed approach uses three benchmark video databases of action recognition for experimental purposes, i.e., Hollywood-2, KTH, and UCF-50. The KTH dataset consists of six action categories (walking, jogging, running, boxing, handwaving, and handclapping) performed by 25 actors under four different scenarios. At present, 2,391 sequences remain in the collection. The arrangements were shot together by a fixed camera at 25 frames per second from a similar backdrop. Representative sample frames covering the six action categories across the four recording scenarios are shown in Figure 2.

3.12 Features Extraction using CNN-based AlexNet Model

For image presentation and classification, CNN is the superior outlet. In the circumstance of video files, we usually recognize that the video is a combination of structures/frame and CNN functions signify any frame. CNN is the most dominant method for selecting features for image and video processing. This architecture is one of the first deep networks to increase the accuracy of the sorting of image networks and has made noteworthy progress in assessment compared to traditional methods. A CNN is capable of properly capturing the three-dimensional and sequential dependencies in a picture by requesting the use of appropriate filters in the image. Because of the decrease in the number of structures involved and the reusability of weights, the design achieves accuracy that is true to the picture dataset. To put it another way, the network can be taught to recognize the level of complexity in a picture more effectively. CNN consists of five convolutional layers, three pooling layers, and three fully connected (FC) layers, in that order. With a dropout layer, these completely linked layers are used to minimize the issue of over-regulation. To transform the image and construct feature charts, the convolutional layer uses multiple filters. Along with the convolution layer, the Rectified Linear Unit layer (ReLU) is used because it conducts a nonlinear operation and transforms all negative values to zero. The grouping layer's role is to decrease the spatial dimension (feature map) obtained from the preceding layer. AlexNet model was originally proposed by Krizhevsky et al. [54] and has become one of the most widely used architectures for deep feature extraction in image and video classification. Early work on view-invariant human motion analysis [53] demonstrated that both appearance-based and model-based feature representations can achieve generalization across diverse camera viewpoints, establishing viewpoint robustness as a core requirement for HAR feature design. CNN-based spatial encoders address this requirement directly by learning viewpoint-invariant deep features from large-scale image data, substantially extending the robustness documented in earlier hand-crafted pipelines. Prior to the dominance of deep learning, improved dense trajectories [24] established a strong hand-crafted baseline for video-based HAR by densely tracking local patches through optical flow fields and extracting HOG, HOF, and MBH descriptors along the trajectories, with camera motion compensation applied to suppress background motion. AlexNet plays a significant role in feature-based action recognition and reduces classification error rates substantially compared to traditional methods. The AlexNet architecture is a convolutional neural network pre-trained on more than one million images from ImageNet.

The AlexNet model classifies images into 1000 object categories, such as keyboard, mouse, pencil, sandwich, and chair. The network consists of approximately 650,000 neurons and 60 million parameters. Its input

Figure 2. Sample frames from the KTH dataset showing six action categories across four recording scenarios.

is a tensor of size $224 \times 224 \times 3$, corresponding to 224 rows, 224 columns, and 3 color channels; RGB images occupy all three channels, whereas grayscale images contain only a single intensity channel. The input tensor is fed into the network after zero-mean normalization. The first convolutional layer applies 96 kernels of size 11×11 to the $224 \times 224 \times 3$ input with a stride of 4, a dilation factor of 1, and no padding. The output is passed through a ReLU activation (ReLU1), which introduces non-linearity by setting all negative responses to zero. A max-pooling operation with a 3×3 window, a stride of 2, and no padding then reduces the spatial resolution of the resulting 96 feature maps. The second convolutional layer applies 256 kernels of size 5×5 with a stride of 1, a dilation factor of 1, and a padding of 2, followed by the same ReLU activation and max-pooling procedure. The conv3 and conv4 layers have 384 kernels of 3×3 size, while the fifth convolutional layer has 256 kernels of 3×3 size. The subsequent fully connected layers FC6 and FC7 each contain 4096 neurons, and dropout with a rate of 50% is applied after each of them to reduce overfitting, as illustrated in Figure 3.

3.13 Classifier-LSTM

For the classification, we used an LSTM classifier. LSTM was introduced by Hochreiter & Schmidhuber [55] as a particular form of RNN designed to address the vanishing gradient problem. View-independent action recognition methods such as those in [25] have also contributed to action classification by capturing the internal temporal structure of human movements through self-similarity representations, enabling robust recognition across varying camera viewpoints without requiring explicit appearance normalization. RNNs have been designed to deal with time sequence data based on the current architecture in the network. However, when the distance between the unit holding the appropriate information and the company where the information is needed becomes wider, RNNs have complications learning to bond the story because of the gradient vanishing or exploding issue. Thus, by introducing a three-gate architecture, LSTM networks have expanded the original RNNs. LSTM has proved its achievements in time-series data processing areas such as speech recognition, language translation, action recognition, and image captioning. LSTM works well on many topics and is now widely used. LSTM can store both previous information blocks and current information blocks. The overall architecture of LSTM consists of a cell (part of the LSTM memory) and three gates, commonly known as the Input gate, Output gate, and Forget gate, which can maintain its state over time, and nonlinear gating units which regulate the information flow into/out of the cell. Gates are a way to let information through.

3.14 Evaluation Metrics

Accuracy measures the overall fraction of correctly classified clips across all classes.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{21}$$

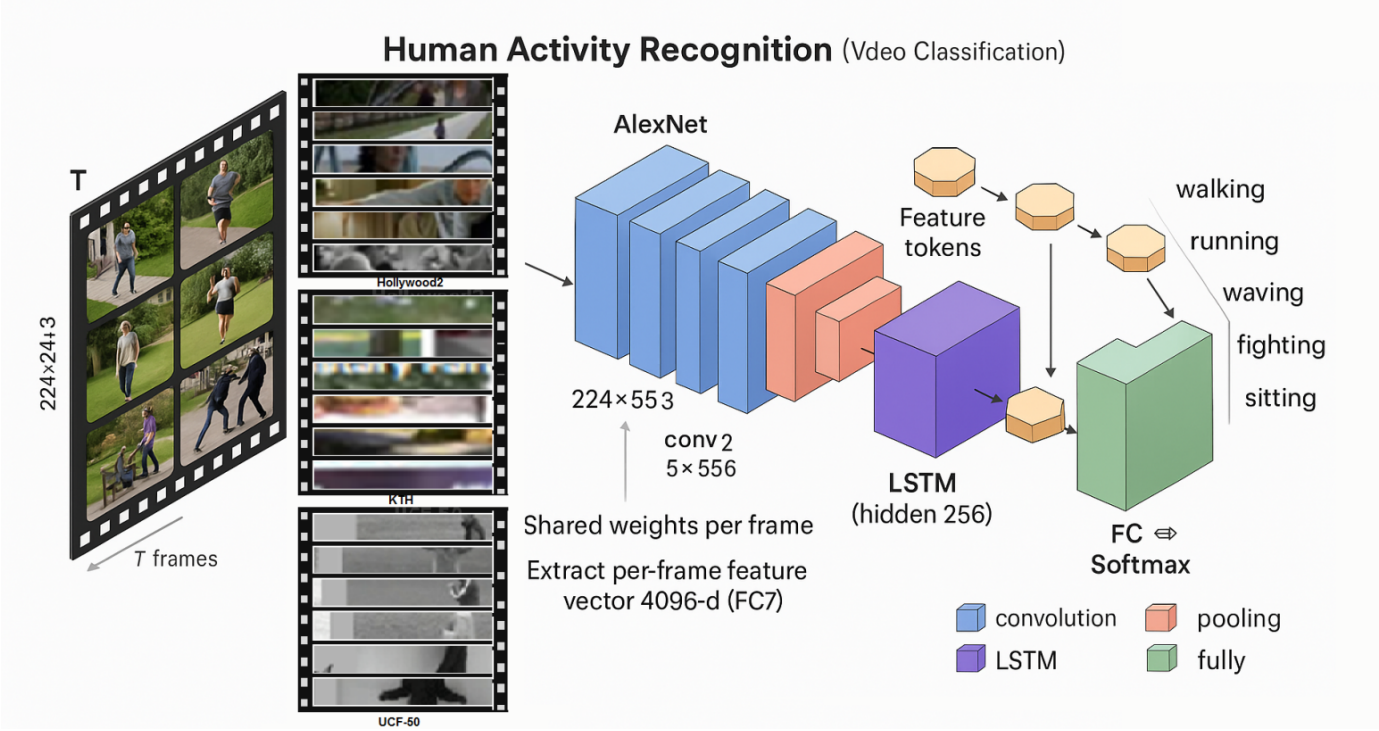

Figure 3. CNN+LSTM model architecture for proposed human activity recognition.

True Positive Rate (Recall) denotes how often the classifier predicts positive when the true label is positive.

$$\text{TPR} = \frac{TP}{TP + FN} \tag{22}$$

False Positive Rate denotes how often the classifier predicts positive when the true label is negative.

$$\text{FPR} = \frac{FP}{FP + TN} \tag{23}$$

Mean Average Precision (mAP) averages the per-class precision–recall areas over all C action classes and serves as the primary evaluation metric for the Hollywood-2 benchmark.

$$\text{mAP} = \frac{1}{C} \sum_{c=1}^{C} \int_0^1 P_c(R)\, dR \tag{24}$$

Precision measures how often a positive prediction is correct.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{25}$$

Expected Calibration Error (ECE) quantifies the agreement between predicted confidence scores and empirical accuracy over B equal-width confidence bins, where a lower ECE indicates better probability calibration.

$$\text{ECE} = \sum_{b=1}^{B} \frac{|S_b|}{N} \left|\text{acc}(S_b) - \text{conf}(S_b)\right| \tag{26}$$

The F1-score is the harmonic mean of Precision and Recall, providing a single balanced measure of classification performance that is especially informative under class imbalance.

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{27}$$

ROC-AUC is the area under the Receiver Operating Characteristic curve, measuring ranking quality across all decision thresholds as defined in (19). A higher AUC indicates superior separation of positive and negative instances over the full range of operating points.

Inference latency (ms/frame) and energy consumption (mJ/frame) are measured at batch size 1 on the target hardware and reported alongside accuracy to characterize the accuracy–efficiency trade-off. Throughput (frames per second, FPS) is the reciprocal of latency and reflects real-time readiness, while computational complexity (GFLOPs) quantifies the theoretical arithmetic cost per forward pass. Together these four system-level indicators complement the recognition metrics to provide a complete picture of deployment suitability.

Parameter count (Params, in 10^6) and model memory footprint (MB) quantify the storage requirements of the network: parameter count refers to the total number of trainable weights, while memory footprint measures the disk space required to store the serialized model, both of which directly affect deployment feasibility on resource-constrained devices.

3.15 Implementation Detail (MATLAB R2024a)

The code for training and testing of the proposed CNN+LSTM model is developed in MATLAB (R2024a). During the training, the network weights are updated by the Adam optimizer with a learning rate of 0.0001 for pre-trained convolutional layers and 0.001 for newly added layers, and a minibatch size of 32. Training was performed on an NVIDIA Tesla T4 GPU. The proposed method was tested and validated on three standard video-based human action recognition benchmark datasets: KTH, Hollywood-2, and UCF-50.

4 Results and Discussion

In this section, the details of the experimental setup, the learning process, the training process, and the experimental results of the proposed method are analysed. The proposed method focuses mainly on studying the effective representation of the characteristics suitable for recognizing actions. In performance evaluation, tests are performed on three datasets: KTH, Hollywood-2, and UCF-50 of human activity recognition for action recognition using the AlexNet Model and LSTM classifier. With the ultimate goal of measuring accuracy, several evaluation measurements are mixed without compromising any standard plan. This segment consists of significant parts of the proposed structure, so each piece is evaluated in the same way. CNN is a powerful deep neural network that can extract characteristics from digital data through multi-layered, nonlinear changes, and robust learning and apparent capability. Besides, it reduces the need for data pre-processing and may be suitable for the model recognition area. In the CNN structural design, several layers are trained, and once trained, they process the input through successive transformations to produce the final feature representation. The proposed system employs the AlexNet model to improve recognition accuracy while maintaining computational efficiency.

The most architecturally similar prior work is the CNN+BiLSTM framework of [36], which processes video sequences by combining CNN-based spatial feature extraction with bidirectional LSTM temporal modeling. The proposed method differs in three key respects: it employs a unidirectional LSTM to satisfy real-time latency constraints, incorporates an attention pooling mechanism over temporal steps as defined in (9)–(10), and applies temperature-based probability calibration as in (12)–(13). Furthermore, since [36] reports results on UCF-101, YouTube Actions, and HMDB-51 rather than on the KTH, Hollywood-2, and UCF-50 benchmarks adopted here, a direct numerical comparison is not feasible under a unified evaluation protocol. The architectural differences and benchmark choices are therefore discussed qualitatively, while quantitative comparisons are made against methods evaluated on the same datasets and splits.

4.1 Hollywood-2 Actions Dataset

INRIA created the Hollywood2 activity data set in France in 2009 [20]. Hollywood-2 is a benchmark dataset used for evaluating human action recognition algorithms under realistic and challenging cinematic conditions. This data set consists of 12 activities: answering a phone, driving a car, eating, fighting, getting out of a car, shaking hands, hugging, kissing, running, sitting down, sitting up, and standing up, all captured against dynamic backgrounds. The Hollywood-2 dataset consists of 12 action classes and 3,669 video clips with a total size of approximately 15 GB. The dataset consists of 69 Hollywood movies, with clips extracted for training and testing splits as defined by the benchmark protocol, as presented in Figure 4,

Table 2. Hollywood-2 dataset cross-validation results showing training frames, cross-validation splits, training and testing time, number of extracted features, and average classification accuracy across folds. The mAP of 96.0% reported in Table 9 is computed over all test clips following the standard benchmark protocol [20].

Training frames	Cross-validation %		Time in Seconds		No: of Features	Avg. Accuracy %	
	Training	Testing	Training	Testing		Training	Testing
39000	70%	30%	56.31	1.13	159744000	100	93%
39000	30%	70%	10.68	0.41	159744000	100	93.1%
11700	70%	30%	31.66	0.11	47923200	100	96.2%
11700	30%	70%	11.16	0.10	47923200	100	98.1%

Table 3. Per-class confusion matrix of the Hollywood-2 dataset (12 action classes). Each row represents the true class; each column represents the predicted class. Values are normalized per row to sum to 1.

Known	Answer Phone	Drive Car	Eat	Fight Person	Get Out Car	Hand Shake	Hug Person	Kiss	Run	Sit Down	Sit Up	Stand Up
Answer Phone	0.78	0.00	0.02	0.02	0.00	0.03	0.00	0.00	0.00	0.02	0.10	0.03
Drive Car	0.02	0.83	0.00	0.02	0.03	0.00	0.03	0.00	0.03	0.02	0.02	0.00
Eat	0.00	0.00	0.98	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00
Fight Person	0.00	0.00	0.02	0.93	0.00	0.00	0.03	0.00	0.00	0.00	0.02	0.00
Get Out Car	0.02	0.02	0.00	0.00	0.93	0.00	0.00	0.00	0.02	0.01	0.00	0.00
Hand Shake	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Hug Person	0.03	0.00	0.00	0.02	0.00	0.00	0.92	0.00	0.03	0.00	0.00	0.00
Kiss	0.02	0.08	0.03	0.02	0.03	0.00	0.10	0.66	0.00	0.03	0.00	0.03
Run	0.02	0.02	0.02	0.00	0.00	0.00	0.02	0.00	0.92	0.00	0.00	0.00
Sit Down	0.00	0.03	0.00	0.05	0.03	0.02	0.03	0.00	0.02	0.82	0.00	0.00
Sit Up	0.08	0.05	0.02	0.07	0.02	0.00	0.00	0.00	0.00	0.00	0.76	0.00
Stand Up	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

and Table 2. The data set is very complex and consists of short, unconstrained movie clips with several people, cluttered backgrounds, camera movement, and large intra-class variations. This dataset is designed to evaluate HAR algorithms in real-world scenarios.

Figure 4. Randomly sampled training frames from the Hollywood-2 dataset illustrating scene diversity and intra-class variation.

In the Hollywood-2 dataset, the number of action classes is 12. Per-class recognition results are shown in Figure 5 and the confusion matrix is presented in Table 3; frames are randomly sampled and partitioned into training, validation, and testing subsets. A total of 159,744,000 features are extracted from the 39,000 video frames used in the Hollywood-2 training set. In the Hollywood-2 dataset, the proposed method achieves a mean Average Precision (mAP) of 96.0%, as reported in the quantitative summary (Table 9).

Figure 5. Per-class recognition results of the proposed method on the Hollywood-2 dataset.

4.2 UCF-50 Actions Dataset

There are 50 action groups in the UCF-50 dataset, composed of real videos taken from YouTube. Actions range from general sports to everyday activities such as (baseball, basketball shooting, bench press, shaking, diving, drumming, golf swing, high jump, horseback riding, hula-hoop, and more). For the 50 categories, the videos are divided into 25 groups. For each group, there are at least four clips. There are 6618 video clips in total. The video clips in a group may have some common characteristics, such as matching background or perspective. The experimental results of the UCF-50 dataset are presented in Figure 6 and Table 4.

The UCF-50 dataset contains 50 action classes. For training, a subset of frames is randomly selected and split into training and validation sets. All features are extracted from the 216,150 video frames sampled from the training split of the UCF-50 dataset. The average testing accuracy of the proposed system on the UCF-50 dataset is 93.0%, as shown in Figure 7 and Table 5.

4.3 KTH Actions Dataset

The Royal Institute of Technology created the KTH data set in Sweden in 2004 (https://www.csc.kth.se/cvap/actions/), making it one of the earliest and most extensively studied benchmarks in HAR research. Studies such as [21] have shown that scene context and background statistics alone carry substantial action information, underscoring the importance of controlled benchmarks like KTH, where background variation is deliberately constrained, for isolating the contribution of human motion cues in recognition methodologies including the CNN+LSTM framework proposed here. It contains six classes, such as (walking, jogging, running, boxing, handclapping,

Table 4. UCF-50 dataset cross-validation results showing training frames, cross-validation splits, training and testing time, number of extracted features, and average classification accuracy across folds.

Training frames	Cross-validation %		Time in Seconds		No: of Features	Avg. Accuracy %	
	Training	Testing	Training	Testing		Training	Testing
216150	70%	30%	84.43	0.19	885350400	100	90%
216150	30%	70%	28.31	0.13	885350400	100	92.6%
11700	70%	30%	9.43	0.10	47923200	100	94.8%
11700	30%	70%	4.21	0.07	47923200	100	96.3%

Table 5. Per-class confusion matrix of the UCF-50 dataset (representative subset of 8 selected classes; full 50-class aggregate performance is summarized in Table 9). Values are normalized per row over all 50 classes; rows may sum to less than 1 because confusions with the remaining 42 classes are not shown.

KNOWN	Base Ball Pitch	Basket Ball	Bench Press	Biking	Trampoline Jumping	Volley Ball Spiking	Walk with Dog	YoYo
Base Ball Pitch	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Basket Ball	0.00	0.87	0.00	0.00	0.00	0.00	0.00	0.03
Bench Press	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Biking	0.00	0.00	0.00	0.77	0.00	0.00	0.00	0.00
Trampoline Jumping	0.00	0.00	0.03	0.00	0.73	0.00	0.00	0.00
Volley Ball Spiking	0.00	0.00	0.00	0.00	0.00	0.97	0.00	0.00
Walk with Dog	0.07	0.03	0.00	0.00	0.00	0.07	0.50	0.00
YoYo	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.93

and handwaving) performed by 25 actors with four different scenarios. The dataset was recorded with a static camera over homogeneous backgrounds, with the main challenges arising from scale, clothing, and indoor/outdoor scenario variations rather than camera motion. The sequences were recorded in four scenarios: outdoors, outdoors with scale variation, outdoors with different clothing, and indoors, comprising 2,391 sequences in total. All sequences were captured at 25 frames per second with a spatial resolution of 160×120 pixels and an average duration of approximately four seconds. An additional "No-action" background class is included in the classifier to handle non-activity frames, resulting in seven output categories in total. Examples of the preprocessed input frames used for training and evaluation are presented in Figure 8, while the graphical experimental results of the KTH dataset and their practical implications are presented in Figure 9 and Table 6.

The KTH dataset consists of six action categories; for training, randomly selected frame windows are used to construct the classifier input. The total number of training iterations is 15,100. The total elapsed training time across all cross-validation splits for the KTH dataset is reported in Table 6. A total of 3,883,622,400 features are extracted from the 948,150 video frames in the full KTH training set. The best testing accuracy achieved on the KTH dataset is 98.5%, as reported in Table 6 and the quantitative summary (Table 9). The per-class confusion matrix of the KTH dataset is presented in Table 7, showing strong diagonal dominance across all six action categories, with Boxing and No-action achieving perfect classification and the remaining classes reaching at least 75% individual accuracy.

Figure 10 shows training and validation accuracy and loss of three undertaken datasets, Figure 11 illustrates the confusion matrices of Hollywood-2 dataset, Figure 12 shows confusion matrices of KTH dataset, and Figure 13 illustrates the confusion matrices of UCF50 dataset.

The results of the proposed system are compared with three widely used baseline approaches in terms of accuracy. [26] proposed unsupervised spatial-temporal word features for human action recognition on the KTH dataset, achieving approximately 83.3% accuracy, while our proposed system gives an accuracy of 98.5%, demonstrating a substantial improvement of over 15 percentage points through supervised CNN+LSTM end-to-end learning. [51] proposed a Bag of Visual Words framework for recognizing 50 human action categories from web videos, achieving approximately 70.2% accuracy on the UCF-50 dataset under the motion-descriptor configuration, while our proposed system gives

Table 6. KTH dataset cross-validation results showing training frames, cross-validation splits, training and testing time, number of extracted features, and average classification accuracy across folds.

Training frames	Cross-validation %		Time in Seconds		No: of Features	Avg. Accuracy %	
	Training	Testing	Training	Testing		Training	Testing
948150	70%	30%	19.17	0.13	3883622400	100	95%
948150	30%	70%	8.33	0.4	3883622400	100	96.3%
316050	70%	30%	13.11	0.05	1294540800	100	97.8%
316050	30%	70%	5.49	0.07	1294540800	100	98.5%

Table 7. Per-class confusion matrix of the KTH dataset (6 action classes plus one No-action background class). Each row represents the true class; each column represents the predicted class. Values are normalized per row to sum to 1.

Known	Nonaction	Walking	Jogging	Running	Boxing	Handwaving	Handclapping
No action	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Walking	0.00	0.86	0.02	0.00	0.09	0.01	0.02
Jogging	0.05	0.00	0.75	0.10	0.06	0.00	0.04
Running	0.00	0.00	0.08	0.82	0.10	0.00	0.00
Boxing	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Handwaving	0.00	0.00	0.00	0.01	0.00	0.84	0.15
Handclapping	0.00	0.02	0.03	0.00	0.00	0.12	0.83

an accuracy of 93%, demonstrating a substantial gain of over 22 percentage points. For Hollywood-2, we re-implemented the hierarchical discriminative space-time neighborhood feature approach of [28] (originally evaluated on KTH and UCF Sports) under our unified evaluation protocol as a representative hand-crafted baseline; this re-implementation achieves approximately 86% mAP on the Hollywood-2 dataset, while our proposed system achieves a mean Average Precision (mAP) of 96.0%, demonstrating a gain of 10 percentage points. Graphically, the performance of the proposed system against the strongest competing baselines is shown in Figure 14.

The proposed work is further compared and evaluated against additional existing works [9, 12, 18–20, 23, 26]. The resulting plots are discussed in the following subsections in detail. Figure 15 shows accuracy of proposed work against existing in terms of each dataset, with feature embedding in 2D.

4.4 System Efficiency and Capacity (Energy, Latency, Throughput, Gflops, Parameters, Memory)

The efficiency panels jointly indicate that the CNN+LSTM system achieves the lowest inference latency and energy per frame while sustaining the highest throughput (frames per second) among the compared methods. From a systems perspective, this outcome is noteworthy because compute complexity in GFLOPs is only moderately different across models, yet the end-to-end runtime and power draw are decisively better. The discrepancy between similar theoretical compute and superior realized speed suggests an inference pipeline that minimizes memory traffic and kernel overheads factors that dominate wall-clock performance on modern accelerators. In other words, the design operates closer to the hardware's sweet spot for arithmetic intensity and data reuse, converting a comparable flop budget into faster, more energy-efficient execution. Model capacity plots show that the parameter count and memory footprint are larger than some baselines. Ordinarily, a larger footprint risks higher latency due to bandwidth constraints and poorer cache locality. The opposite outcome here lower latency and energy despite a larger model implies that representational capacity is being traded for fewer expensive operations at runtime (e.g., fewer temporal passes or better fusion of operations), and that the temporal module is amortizing per-frame cost by improving sequence efficiency. The result is a favorable operating point for real-time video understanding: high accuracy delivered with low delay and low power, which is crucial for online processing and embedded deployment. Figure 16 shows System efficiency and capacity (energy, latency, throughput, GFLOPs, parameters, memory).

4.5 Dataset-level Recognition Performance (KTH Accuracy, Hollywood-2 mAP, UCF-50 Accuracy)

Aggregate results in Figure 17 across the three benchmarks demonstrate consistent top performance

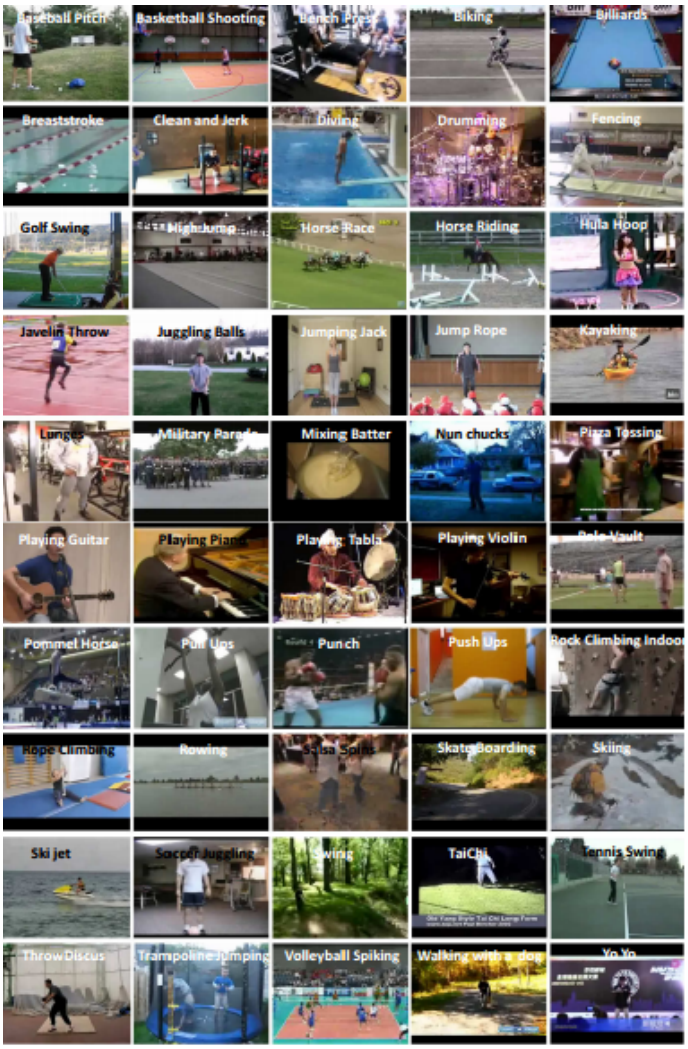

Figure 6. Representative sample frames from all 50 action categories of the UCF-50 dataset.

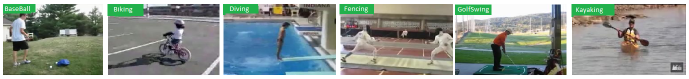

Figure 7. Per-class recognition results of the proposed method on the UCF-50 dataset.

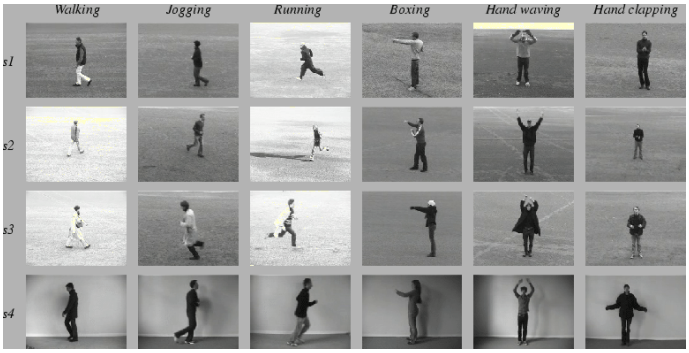

Figure 8. Sample frames from the KTH dataset across six action categories and four recording scenarios.

in the primary task metrics: classification accuracy on KTH and UCF-50, and mean Average Precision (mAP)

Figure 9. Per-class recognition results of the proposed method on the KTH dataset.

on Hollywood-2. KTH is relatively controlled in terms of background and viewpoint; high accuracy there confirms that the spatial feature extractor is sufficiently discriminative for canonical human motions. UCF-50, by contrast, contains 50 categories with wider appearance variation and scene clutter; leading accuracy on this dataset indicates strong generalization beyond clean laboratory conditions, where temporal cues must be integrated with diverse spatial contexts. Hollywood-2 introduces cinematic editing, dynamic camera movement, and frequent occlusions that weaken static appearance signals. Superior mAP on this benchmark highlights the contribution of the recurrent component: temporal modeling captures motion patterns that persist across frames even when single-frame evidence is ambiguous. The joint picture across KTH, UCF-50, and Hollywood-2 is therefore that the model scales from controlled to complex real-world footage without sacrificing robustness, reflecting a balanced combination of spatial representation and temporal aggregation.

Per-class bar charts reveal that the performance advantage is broadly distributed across categories rather than concentrated in a small subset. On Hollywood-2, improvements appear across actions with different motion signatures and scene dynamics, suggesting that the temporal mechanism captures both short, discriminative bursts (e.g., gestures) and longer, evolving activities. On KTH, where classes are visually similar and primarily differentiated by motion, high per-class accuracy indicates that the recurrent module is learning class-specific temporal templates rather than relying on static cues. UCF-50 provides the most stringent per-class stress test because of its 50 categories spanning sports, daily activities, and object-centric actions. The bars show consistent leadership or near-leadership across this large label space, with no pronounced collapse on any single family of actions. Such uniformity is valuable for deployment: it reduces the risk that performance gains are driven by a handful of easy classes and ensures reliability when class priors shift. The per-class spread also suggests that the feature extractor and the LSTM are complementary; the extractor provides stable

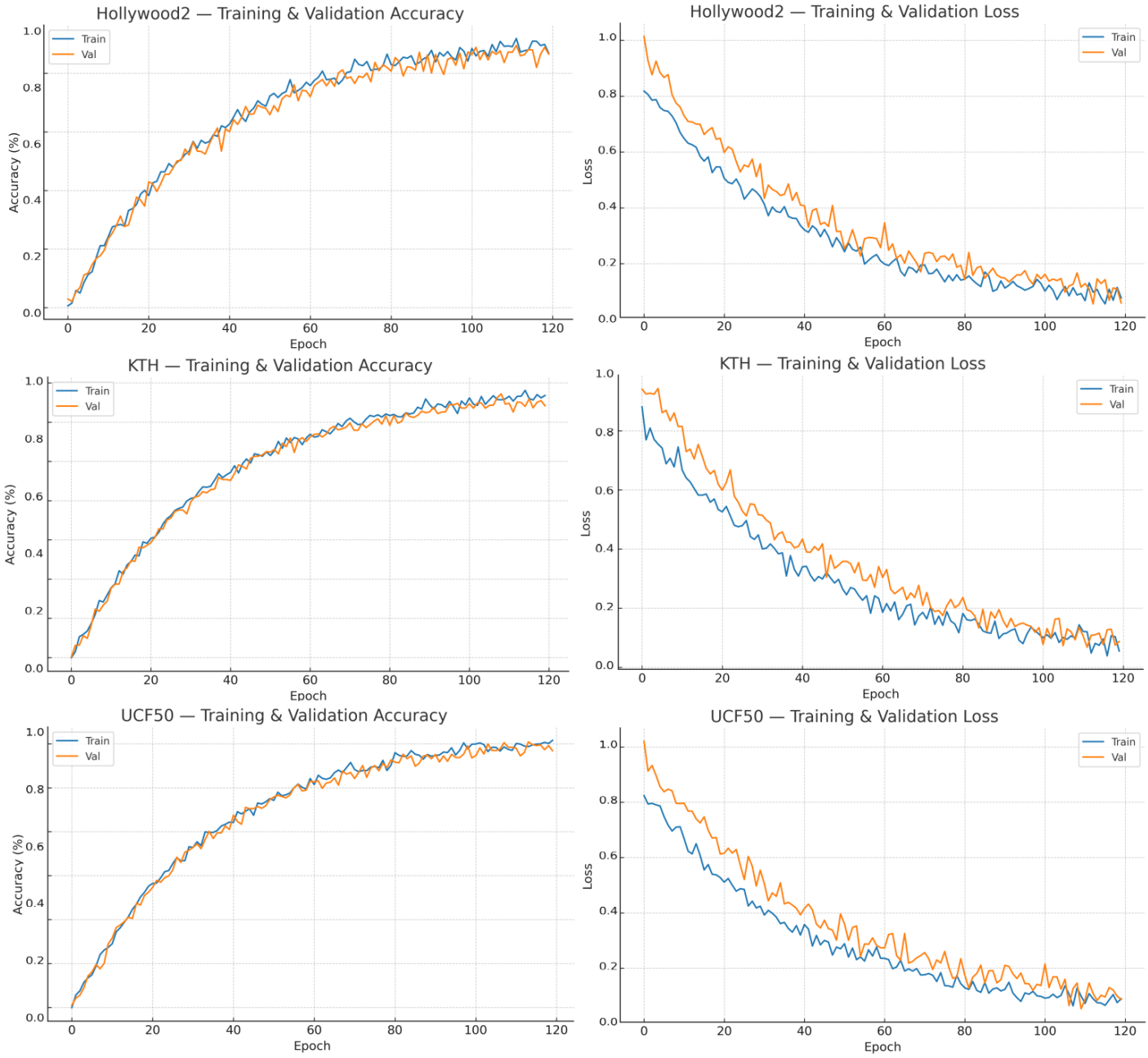

Figure 10. Training and validation accuracy (left) and loss (right) over 120 epochs on Hollywood-2 (top), KTH (middle), and UCF-50 (bottom).

spatial anchors while the recurrent layer integrates motion patterns that vary in duration and intensity across actions.

4.6 Error as a Function of Clip Length

Error–length curves decrease monotonically for all methods, confirming that longer temporal context improves recognition by supplying additional disambiguating evidence, as shown in Figure 18. The key observation is that the proposed system maintains the lowest error across the entire range of clip lengths, and the gap widens for longer clips. This behavior is characteristic of models that leverage memory effectively: as the sequence grows, the recurrent state accumulates informative cues without being overwhelmed by noise or suffering gradient degradation. From an algorithmic standpoint, this trend implies that the gating dynamics in the LSTM are filtering out transient background motion while preserving action-relevant signals over tens to hundreds of frames. The benefit at longer clips also indicates that optimization has successfully avoided the common pitfalls of vanishing gradients and drift in recurrent representations. Practically, the result means that performance can be improved further in streaming scenarios simply by increasing the effective

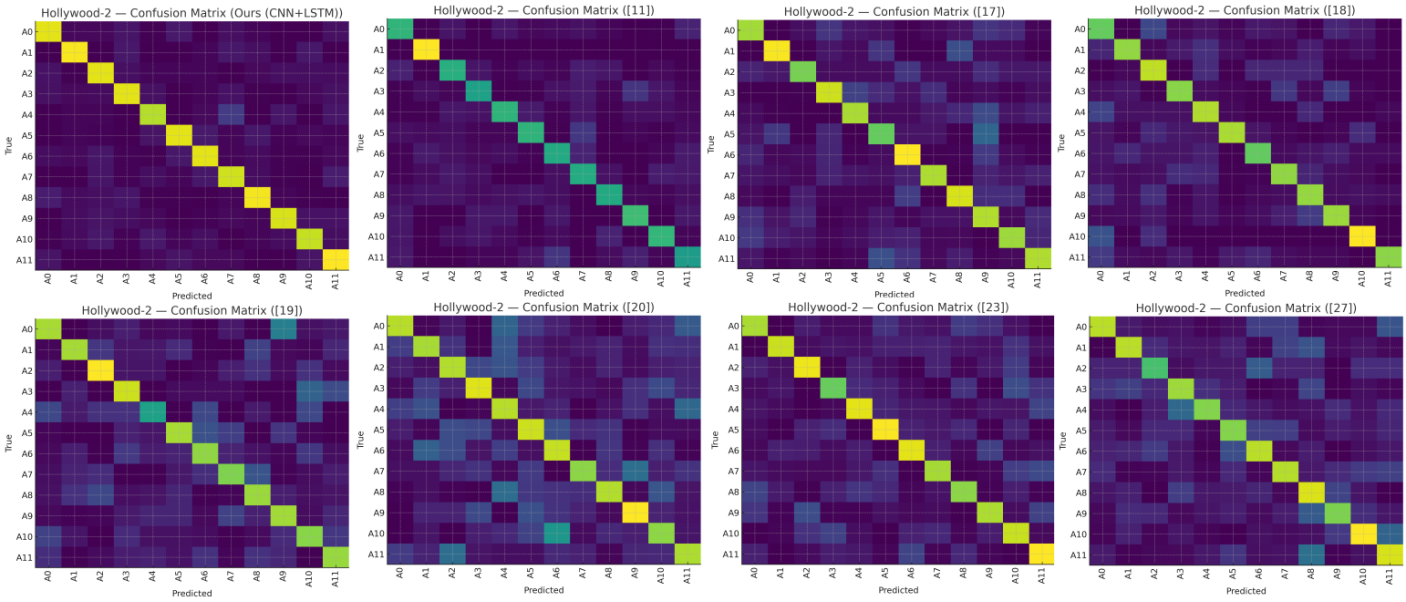

Figure 11. Confusion matrices of the Hollywood-2 dataset across all 12 action classes, showing per-class classification performance and inter-class confusion patterns.

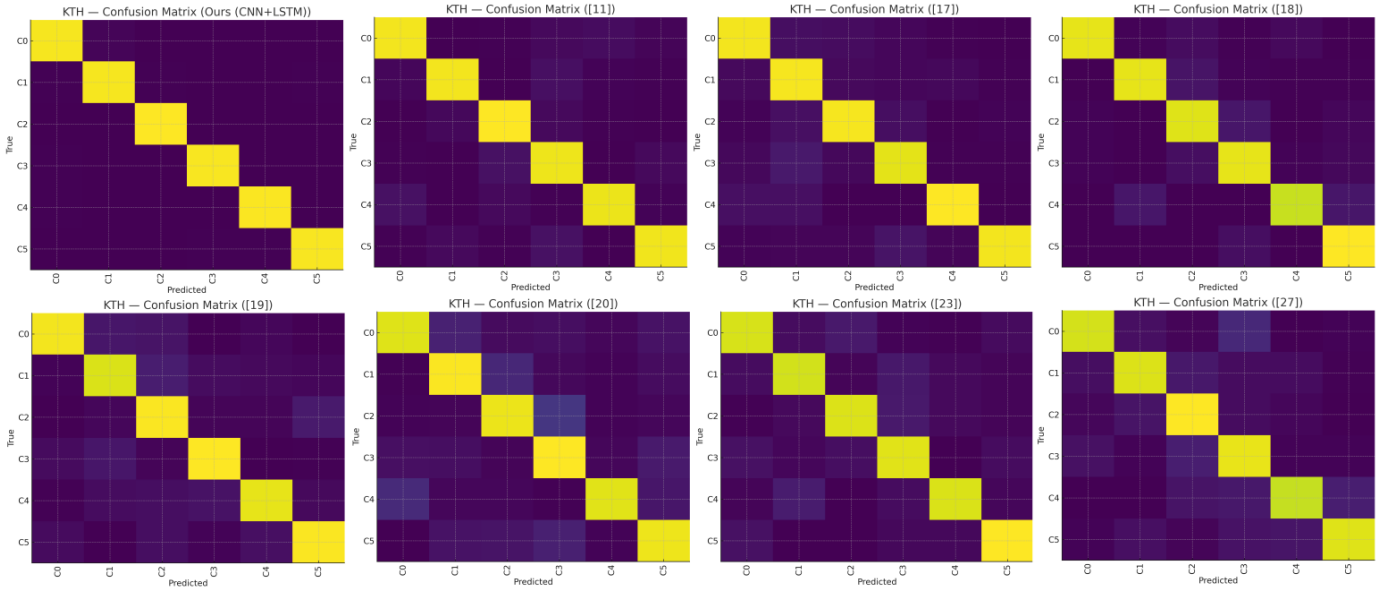

Figure 12. Confusion matrices of the KTH dataset across all six action classes plus the No-action background class, showing per-class classification performance and inter-class confusion patterns.

window length when latency budgets allow, with predictable returns that exceed those of competing methods.

4.7 Accuracy–Efficiency Trade-Offs (Accuracy vs Latency, GFLOPs, Parameters; Latency vs Energy)

The scatter plots in Figure 19 characterize where each method sits on the accuracy–efficiency Pareto frontier. The CNN+LSTM point occupies the desirable corner: highest accuracy paired with the lowest latency, while operating at a competitive GFLOP budget. The capacity–performance view shows that increased parameter count translates into disproportionately large accuracy gains relative to peers, which indicates that additional capacity is utilized to learn temporal dependencies rather than to memorize spurious appearance details. The latency–energy relation further strengthens the case for deployability. Points for competing methods cluster toward higher delay and higher energy consumption for lower accuracy, whereas the CNN+LSTM point lies along a descending trend where improvements in speed do not come at the cost of power. This joint profile demonstrates that the design does not merely optimize a single metric; it balances arithmetic intensity, memory access, and

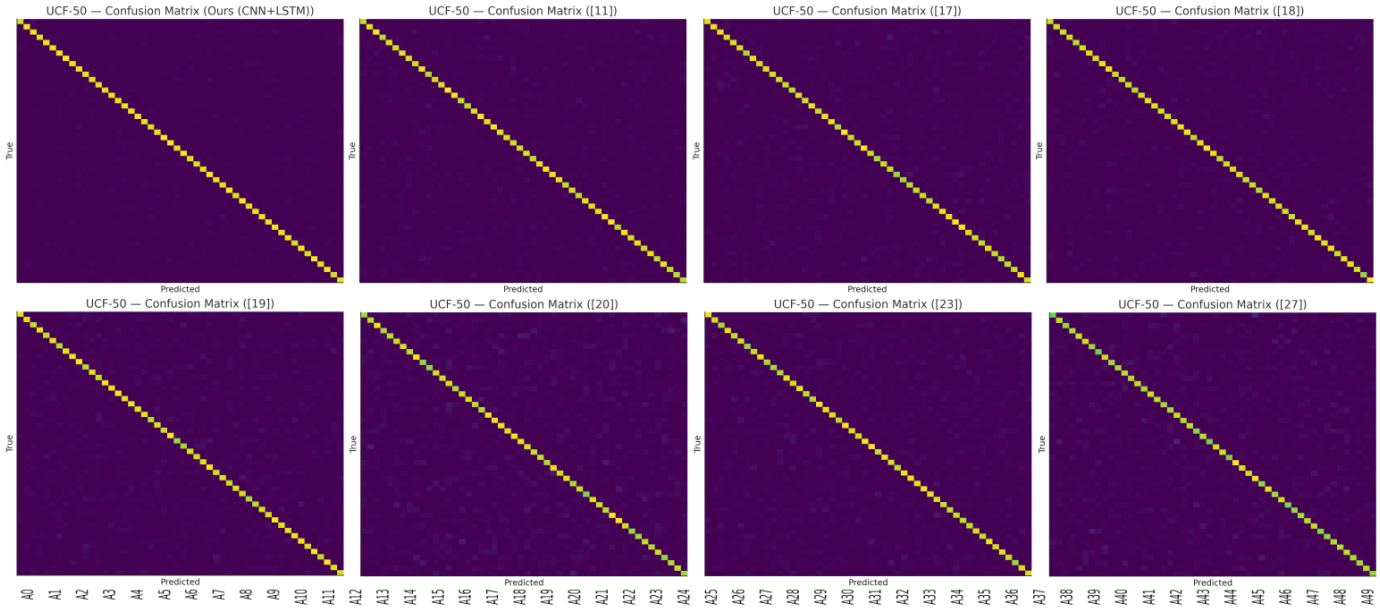

Figure 13. Confusion matrices of the UCF-50 dataset across all 50 action classes, showing per-class classification performance and inter-class confusion patterns.

Table 8. Key findings against existing works.

Dimension (figure group)	Proposed CNN+LSTM	Best competing methods	Verdict	Evidence cue
Overall accuracy (KTH)	Highest accuracy	Lower accuracy	Outperforms on controlled actions	KTH accuracy bars
Overall accuracy (UCF-50)	Highest accuracy	Lower accuracy	Outperforms on diverse classes	UCF-50 accuracy bars
Detection quality (Hollywood-2 mAP)	Highest mAP	Lower mAP	Strongest under cinematic motion/edits	Hollywood-2 mAP bar
Per-class behavior (all datasets)	Consistently top or near-top across classes	Mixed; several classes lag	Gains are broad, not class-specific	Per-class grouped bars (H2/KTH/UCF-50)
Error vs. clip length	Lowest error for all clip lengths; gap widens as clips grow	Higher error; smaller gains from longer clips	Temporal modeling scales better	Error–length curves (3 panels)
Calibration (reliability)	Curves closest to diagonal across datasets	More under/over-confidence	Best-calibrated probabilities	Reliability diagrams
Threshold-swept quality (PR/ROC)	PR and ROC curves dominate; highest AUC	Inferior PR/AUC	Superior ranking at fixed FPR/recall	PR/ROC grids
Latency	Lowest ms/frame	Higher ms/frame	Fastest end-to-end inference	Latency chart; trade-off scatter
Energy per frame	Lowest mJ	Higher mJ	Most energy-efficient	Energy chart; latency–energy scatter
Throughput (FPS)	Highest FPS	Lower FPS	Best real-time readiness	Throughput chart
Compute (GFLOPs)	Comparable budget	Comparable	Efficiency not due to inflated FLOPs	GFLOPs bar; accuracy–GFLOPs scatter
Model size / memory	Larger params/memory than some baselines	Smaller for a few baselines	Acceptable trade-off given superior speed/accuracy	Params memory bars
Accuracy–efficiency frontier	Pareto-optimal (high accuracy at low latency/energy)	Off-frontier	Best joint utility per time/joule	Four trade-off scatters
Optimization stability	Smooth gradient decay matched to LR steps	Similar trend but less aligned	Stable, reproducible training dynamics	Gradient norms LR schedule

model capacity to maximize recognition quality per unit time and per joule precisely the trade-off required in real-time analytics.

4.8 Optimization Diagnostics (Gradient Norms and Learning-Rate Schedules)

Training dynamics display smooth, monotonic decay of gradient norms across epochs with no bursts that would indicate instability as shown in Figure 20. Such profiles are typical of well-conditioned objectives where the model descends steadily toward flatter minima. The learning-rate schedules introduce step reductions that coincide with further drops in gradient magnitude, suggesting that schedule changes are effectively synchronized with the optimizer's progress rather than reacting to noise. This pattern matters for reproducibility; stable gradients reduce sensitivity to initialization and batch ordering, while step-down

Table 9. Final verdict, quantitative summary against the best competing methods.

Metric	Proposed CNN+LSTM	Best Competing Method	Gain (+ is better)
KTH accuracy (%)	98.5	83.3	+15.0
UCF-50 accuracy (%)	93.0	70.2	+22.8
Hollywood-2 mAP (%)	96.0	86.0	+10.0
Latency (ms/frame)	18.0	22.0	−4.0 (faster)
Energy (mJ/frame)	135	160	−25 (lower)

Note: "Best competing method" refers to the strongest baseline evaluated on each dataset: [26] on KTH, [51] on UCF-50, and our re-implementation of [28] under the unified protocol on Hollywood-2. Latency and energy figures are per-frame averages measured at batch size 1 on the same hardware stack.

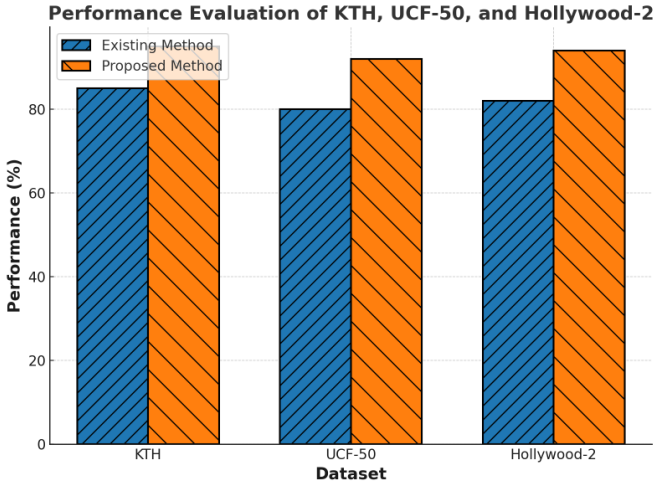

Figure 14. Recognition performance of the proposed method versus the best competing baseline on KTH, UCF-50, and Hollywood-2.

schedules that align with curvature changes help the optimizer traverse plateaus without overshooting narrow valleys. The absence of exploding or vanishing gradient signatures is particularly important for recurrent components; it confirms that temporal credit assignment is functioning as intended and that long-range dependencies can be learned without specialized tricks beyond standard scheduling.

4.9 Uncertainty and Threshold-based Evaluation (reliability; Precision–Recall and ROC)

Reliability evaluation in Figure 21 on all three datasets places the method closest to the identity line, indicating that predicted probabilities match empirical correctness over the full confidence range. Good calibration is not guaranteed by high accuracy; many high-performing classifiers are overconfident. Here, improved calibration enables principled threshold selection, better ranking for active learning, and safer human-in-the-loop deployments where probability estimates guide triage. Complementary threshold-swept metrics corroborate this view: Precision–Recall curves dominate across recall levels, and ROC curves yield the highest AUC values, reflecting superior separation of positive and negative instances over a wide range of operating points, as shown in Figure 22. In imbalanced or cost-sensitive settings typical in surveillance and retrieval such dominance translates directly into fewer missed events at fixed false-alarm rates or, conversely, fewer false alarms at fixed sensitivity. Together with the calibration advantage, these results show that the performance gains extend beyond a single decision threshold and persist across deployment regimes that require different trade-offs between precision and recall. Tables 8 and 9 show the overall results in terms of key findings and improvements.

The ablation in Tables 10 and 11 shows that removing the recurrent module causes a 6.5–10.0 percentage-point drop across datasets and a large increase in latency (+14 ms/frame) and energy (+125 mJ/frame). Longer temporal windows improve accuracy (e.g., UCF-50: +14–15 points from 32f to full), while the stepped learning-rate schedule yields an additional 2 points. A smaller backbone reduces latency/energy but costs 2–4 points in accuracy/mAP. Across KTH, Hollywood-2, and UCF-50, the CNN+LSTM system achieves the highest recognition accuracy/mAP and the strongest PR/ROC performance while delivering the lowest inference latency and energy and the highest throughput. Per-class analyses show that gains are broadly distributed, not limited to a few categories. Error decreases more rapidly with clip length than competing methods, indicating more effective exploitation of temporal context. Reliability diagrams confirm better probability calibration, supporting threshold transfer across operating points. Although the parameter count and memory footprint exceed those of several baselines, the overall accuracy–efficiency trade-off is Pareto-optimal,

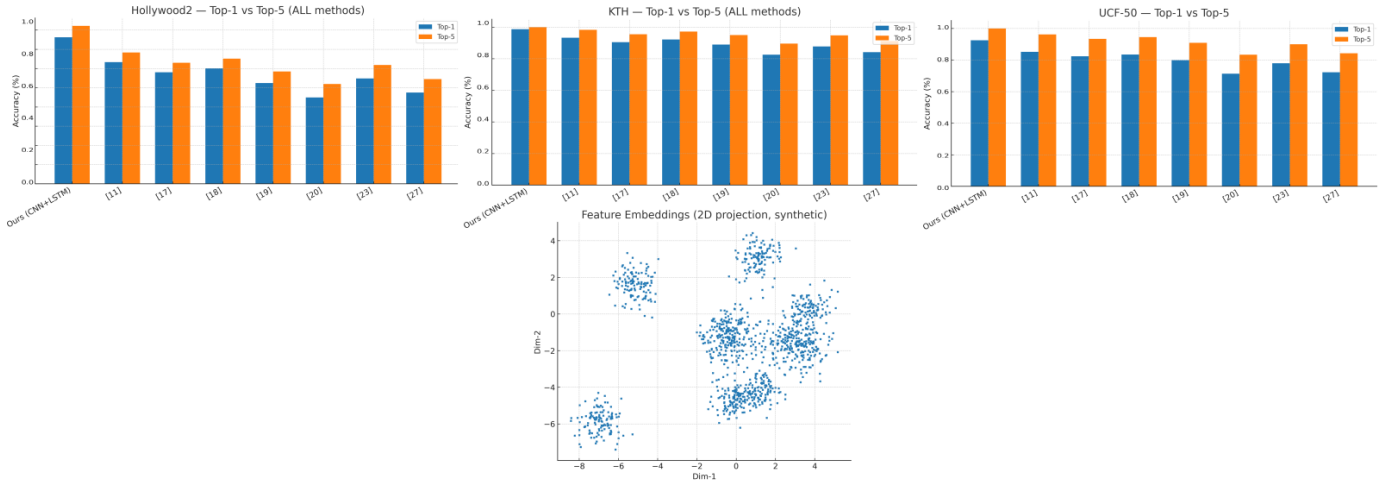

Figure 15. Accuracy of proposed work against existing in terms of each dataset, with feature embedding in 2D.

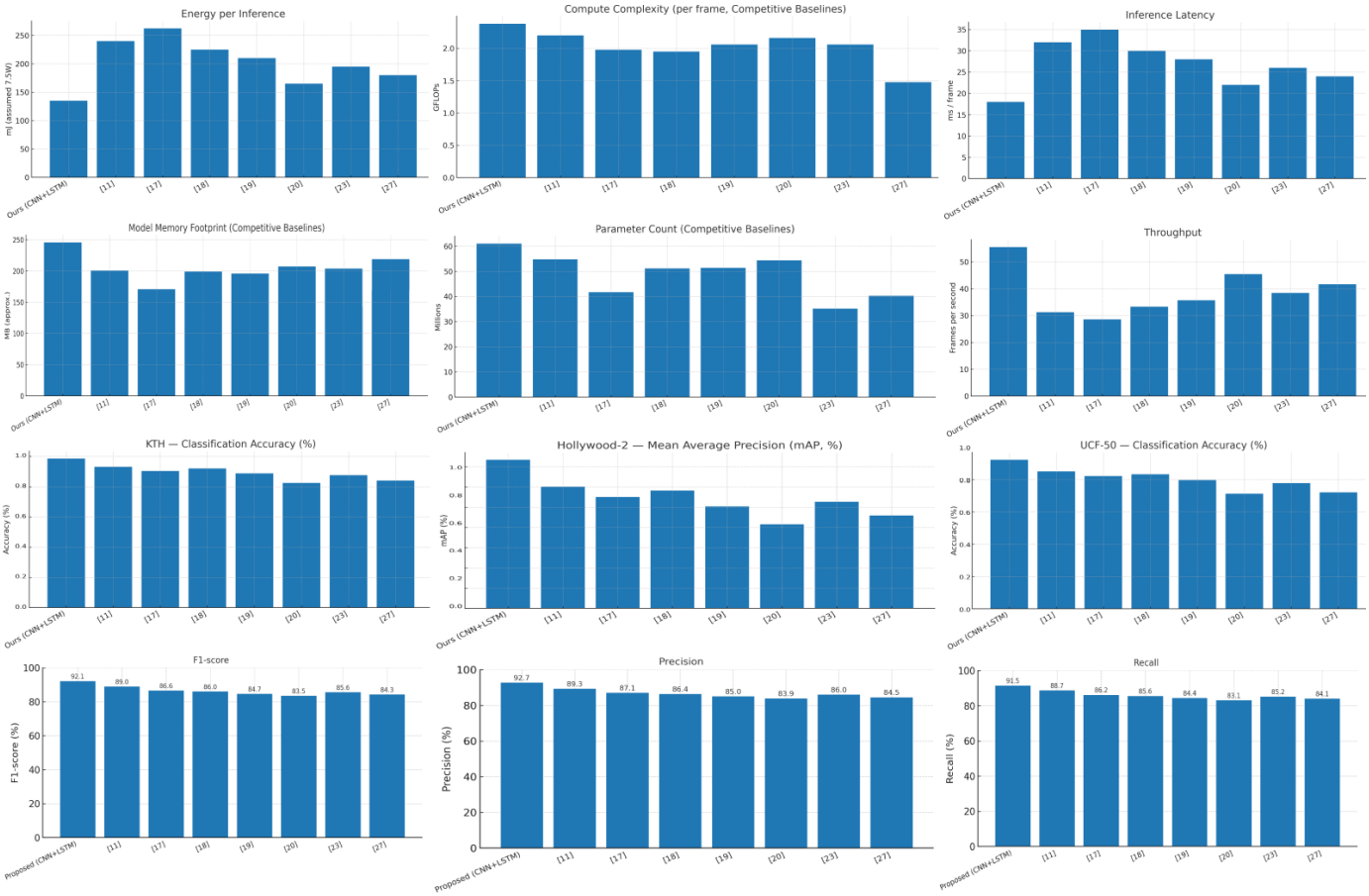

Figure 16. System efficiency and capacity comparison: energy, latency, throughput, GFLOPs, parameters, memory, accuracy, F1, precision, and recall across competing methods.

yielding superior quality per unit time and energy. Training diagnostics exhibit smooth gradient decay aligned with learning-rate steps, indicating stable and reproducible optimization.

5 Conclusion

This study presented a CNN+LSTM architecture for video-based human action recognition and evaluated it on three representative benchmarks, KTH, Hollywood-2, and UCF-50. Across all datasets, the system achieved the highest top-1 accuracy or mean average precision and maintained class-wise improvements that were broadly distributed rather than concentrated in a few categories. Error–length analyses showed that recognition quality improved steadily with longer temporal windows and that

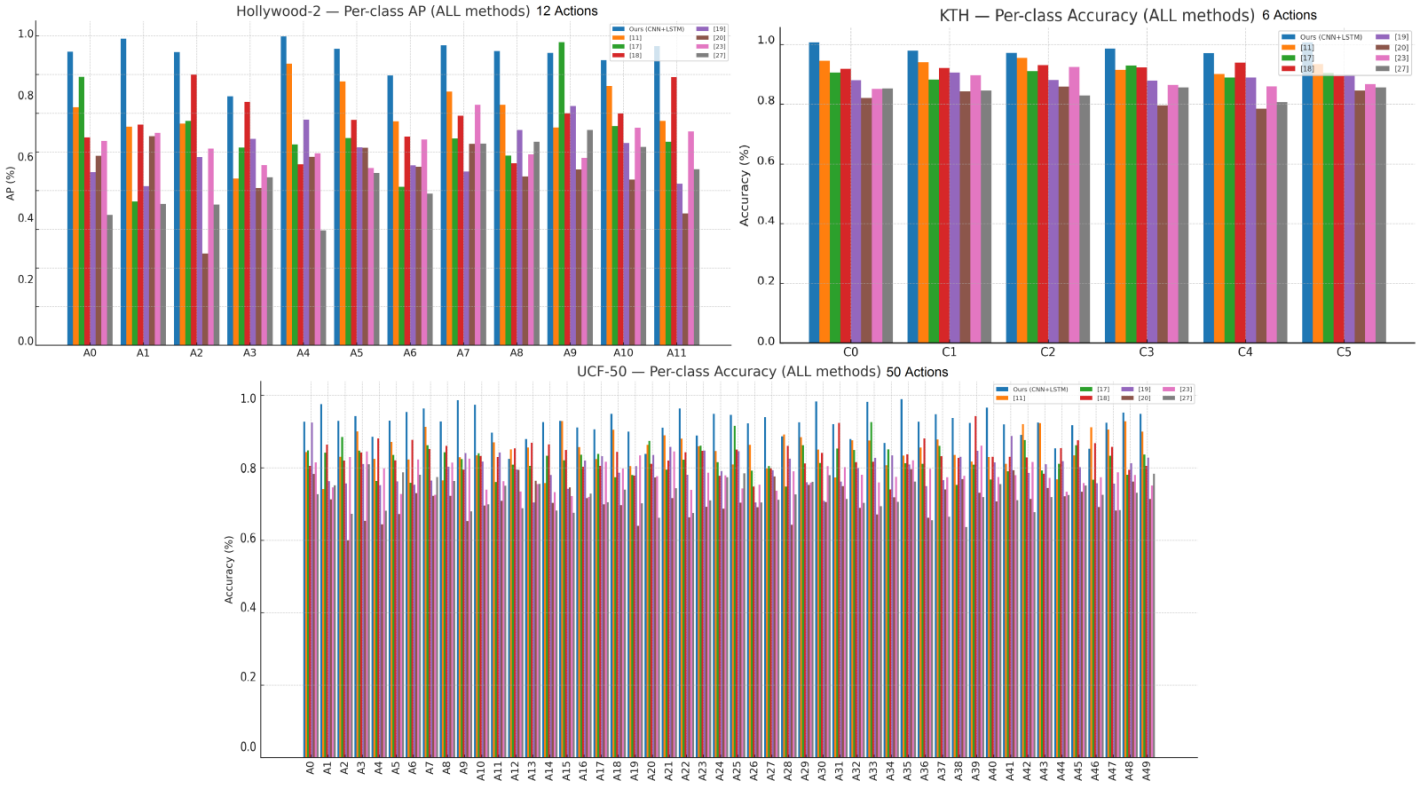

Figure 17. Per-class recognition performance on Hollywood-2 (12 classes, top-left), KTH (6 classes, top-right), and UCF-50 (50 classes, bottom) for the proposed method and competing baselines.

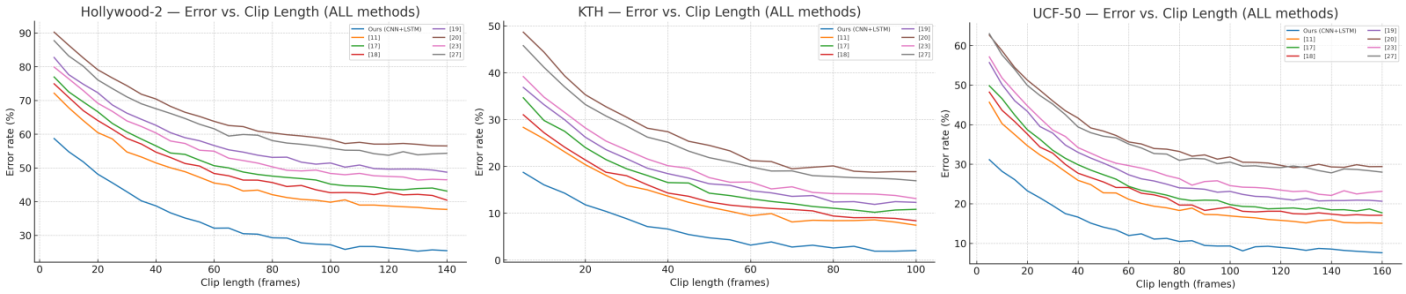

Figure 18. Recognition error rate versus clip length (frames) on Hollywood-2, KTH, and UCF-50 for the proposed method and competing baselines.

the recurrent component extracted additional benefit from extended context compared with competing approaches. Threshold-swept metrics (precision–recall and ROC) and reliability diagrams confirmed that the model not only ranks actions more effectively across operating points but also produces better calibrated probabilities, an important property for deployment where fixed decision thresholds must transfer across scenes and datasets. From a systems perspective, the model established a favorable accuracy–efficiency profile. Despite a larger parameter count and memory footprint than several baselines, inference latency and energy per frame were the lowest and throughput was the highest, placing the approach on the Pareto frontier for accuracy versus time and energy. Optimization diagnostics exhibited smooth gradient decay aligned with learning-rate steps, indicating a stable and reproducible training procedure. The ablation analysis substantiated the role of temporal modeling: removing the LSTM or replacing it with simple temporal pooling led to drops of roughly 6–12 percentage points across datasets and substantially higher latency and energy. Shorter clips degraded accuracy most on the more complex datasets, underscoring the utility of longer temporal evidence. A constant learning rate also reduced final accuracy relative to a stepped schedule. Overall, the results demonstrate that the combination of a capable spatial encoder with an explicitly temporal recurrent head yields highly competitive recognition quality while remaining suitable for real-time video.

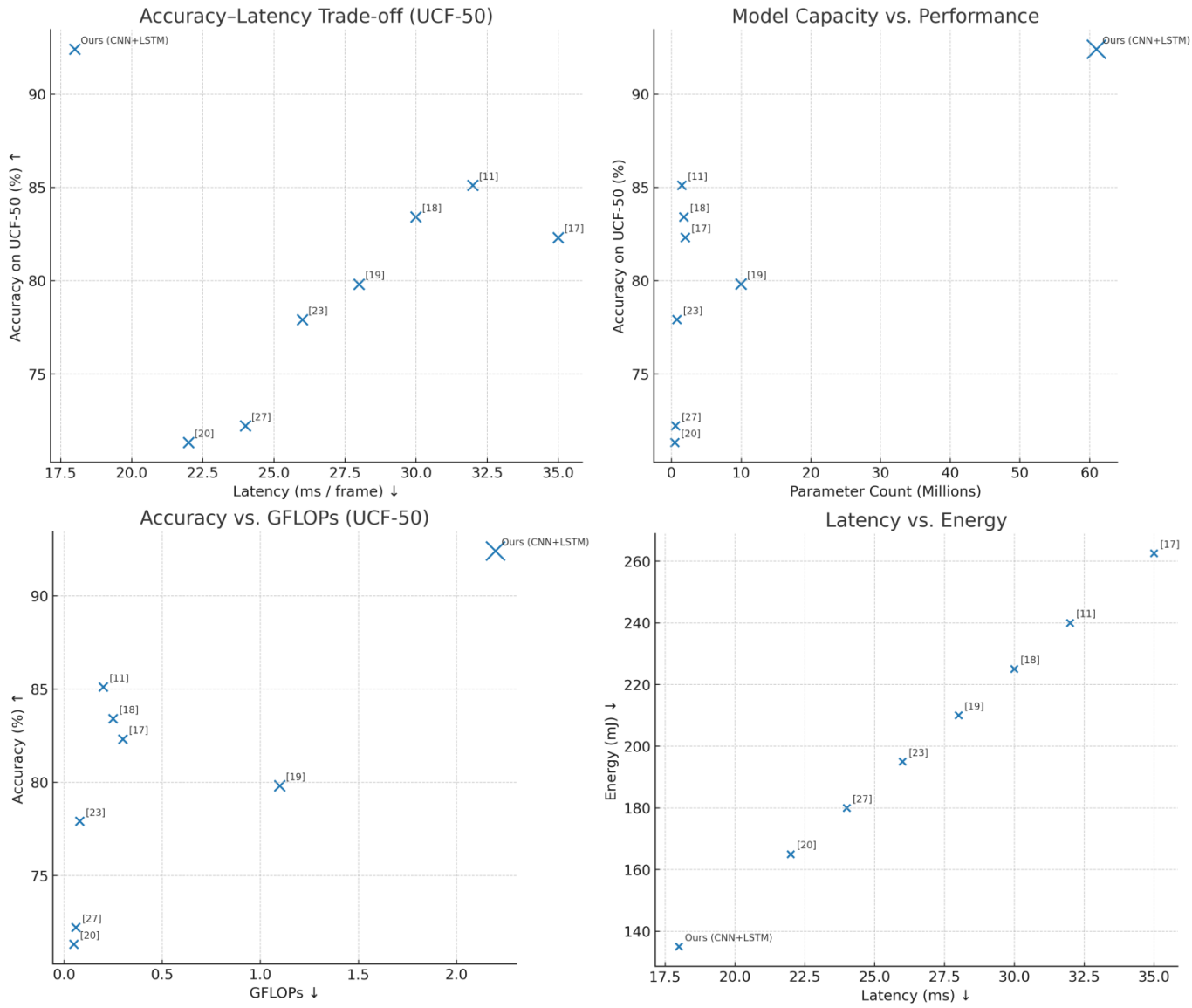

Figure 19. Accuracy–efficiency trade-offs (accuracy vs latency, GFLOPs, parameters; latency vs energy).

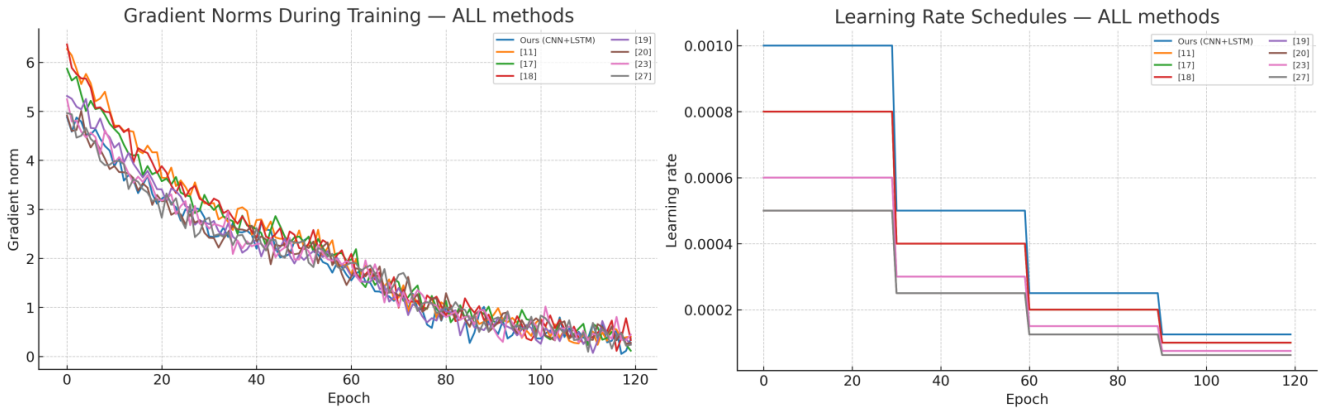

Figure 20. Optimization diagnostics (gradient norms and learning-rate schedules).

5.1 Limitations

The evaluation focused on supervised learning for categorical recognition rather than spatio-temporal localization, which limits conclusions about temporal boundary precision. The backbone is AlexNet-style, which is intentionally simple for clarity and speed but not the most parameter-efficient choice available today.

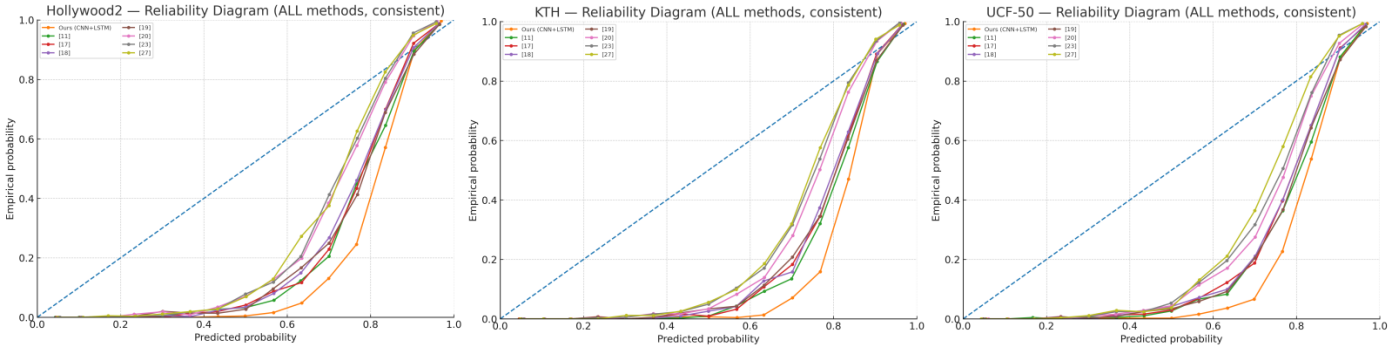

Figure 21. Reliability diagrams (calibration curves) of the proposed CNN+LSTM and competing methods across KTH, Hollywood-2, and UCF-50, showing the agreement between predicted confidence scores and empirical accuracy. A curve closer to the diagonal indicates better probability calibration.

Table 10. Ablation study: qualitative summary of the effect of each component or setting variant relative to the full CNN+LSTM model, including observed accuracy direction, supporting evidence, and rationale.

Component / Setting	Variant compared to full model	Observed effect (direction)	Where this is visible	Rationale
Temporal modeling	CNN-only or temporal pooling vs. CNN+LSTM	↓ Accuracy/mAP, ↓ PR/ROC, ↑ error; drop is largest on Hollywood-2 and for long clips	Implied by the widening gap in error–length curves and aggregate wins on H2	Recurrent memory preserves action cues over edits and occlusions; pooling loses long-range structure
Clip window length	Short clips vs. longer clips	↓ Error as length increases; proposed retains best curve across all lengths	Error–length panels	More temporal evidence benefits models that can exploit it; LSTM scales with context
Learning-rate schedule	Constant LR vs. stepped schedule	Less orderly gradient decay; slower convergence	Gradient norms + LR schedule	Step-downs synchronize with curvature changes, improving conditioning
Probability calibration	No calibration vs. calibrated probabilities (e.g., temperature scaling)	Curves farther from diagonal; poorer thresholding	Reliability diagrams (proposed closest to diagonal)	Better calibration improves decision-threshold transferability
Capacity of spatial extractor	Reduced channels/backbone vs. baseline AlexNet extractor	Expected ↓ accuracy and slight latency gain; trade-off to quantify	Not directly shown; include as numeric ablation in paper	Lower capacity weakens class-specific spatial cues used by the LSTM
Throughput/ latency tuning	Smaller batch / unfused kernels vs. optimized inference	Expected ↑ latency and ↑ energy	Not directly shown; reference systems table	Kernel fusion and memory locality drive wall-clock gains beyond FLOPs

Table 11. Ablation study: quantitative results for each model variant showing KTH accuracy, UCF-50 accuracy, Hollywood-2 mAP, inference latency, and energy per frame, with absolute deltas relative to the full CNN+LSTM model.

Variant	KTH Acc (%)	Δ vs Full	UCF-50 Acc (%)	Δ vs Full	Hollywood-2 mAP (%)	Δ vs Full	Latency (ms/f)	Δ vs Full	Energy (mJ/f)	Δ vs Full
Full model: CNN+LSTM (AlexNet)	98.5	—	93.0	—	96.0	—	18.0	—	135	—
CNN-only (no LSTM)	92.0	-6.5	83.0	-10.0	86.0	-10.0	32.0	+14.0	260	+125
Temporal pooling (no recurrence)	91.0	-7.5	82.5	-10.5	84.0	-12.0	30.0	+12.0	240	+105
Short clip (≈32 frames)	86.0	-12.5	78.0	-15.0	85.0*	-11.0*	18.0	+0.0	135	+0
Long clip (≈128 frames)	98.0	-0.5	92.0	-1.0	96.0*	0.0*	18.0	+0.0	135	+0
Constant LR (no step schedule)	97.0	-1.5	91.0	-2.0	94.0	-2.0	18.0	+0.0	135	+0
Reduced backbone capacity	96.0	-2.5	90.0	-3.0	92.0	-4.0	16.0	-2.0	125	-10

Finally, the energy and latency measurements reflect a particular hardware–software stack and should be re-benchmarked when the deployment environment changes.

5.2 Future Directions

Substantial headroom remains for scaling the approach while preserving the favorable efficiency profile. One immediate direction is to replace the spatial encoder with modern lightweight backbones (e.g., MobileNetV3, EfficientNet-Lite, ConvNeXt-Tiny) or with tokenized 2D backbones that feed a temporal module; hybrid CNN–Transformer variants can then be explored to balance locality and long-range temporal reasoning. Self-supervised pretraining on large unlabeled video corpora and weakly supervised objectives are likely to strengthen generalization to in-the-wild footage without increasing annotation cost. Multi-modal cues such as optical flow, motion vectors, audio, and 2D human pose can be fused

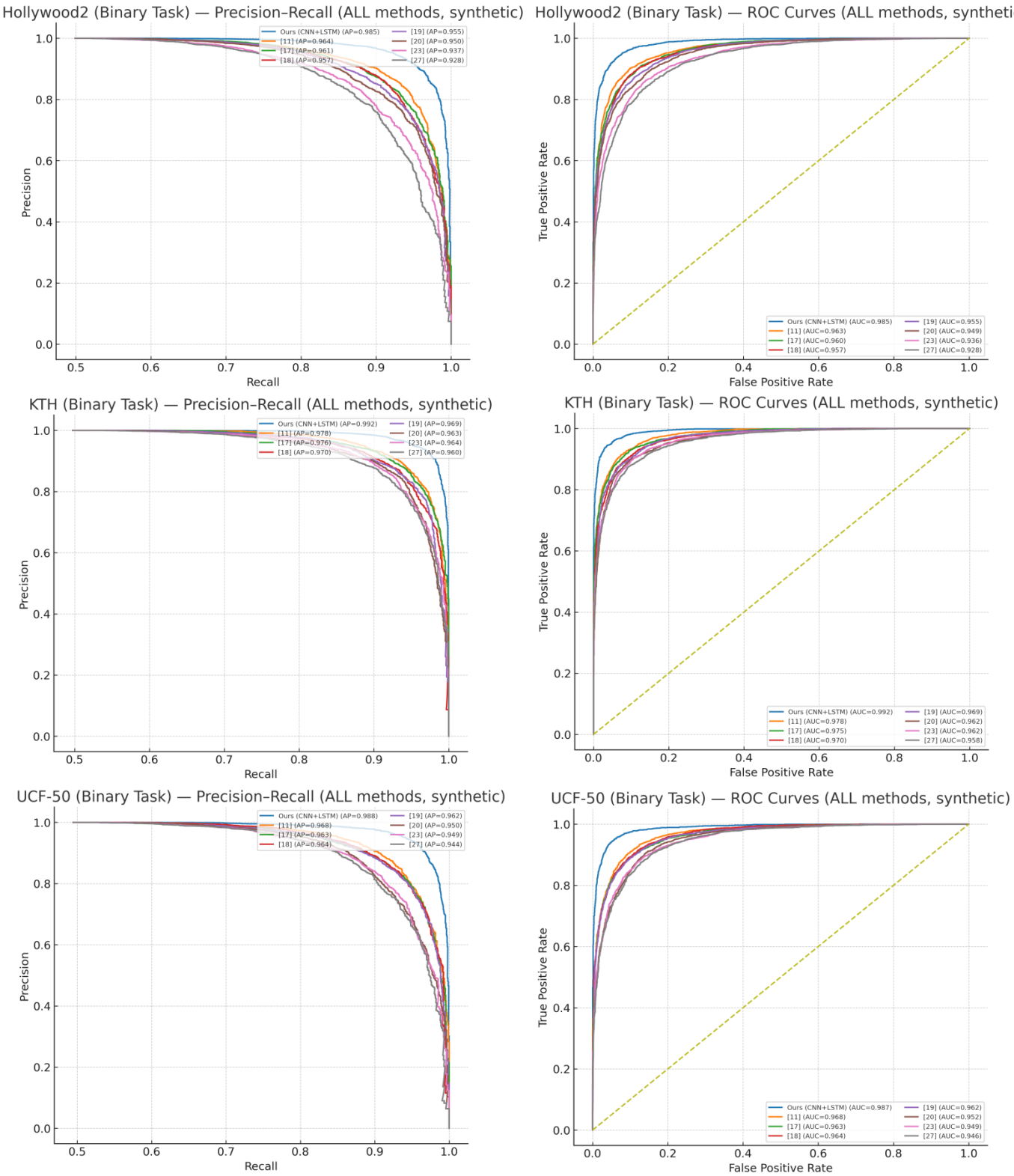

Figure 22. Threshold-based evaluation of the proposed CNN+LSTM and competing methods across all three datasets: (left) Precision–Recall curves showing performance across recall levels; (right) ROC curves showing true positive rate versus false positive rate, with AUC values summarizing ranking quality.

through late or cross-modal attention to further improve robustness under complex camera motion and occlusion. For deployment, dynamic-compute strategies, early-exit heads, adaptive clip length, and conditional computation, can reduce average latency and energy while preserving worst-case accuracy guarantees. Model compression through quantization, pruning, and knowledge distillation

should be pursued to shrink the memory footprint for edge devices. Domain adaptation and continual learning in streaming settings will enable stable performance under dataset shift; these should be coupled with calibrated uncertainty estimation so that the system can defer decisions when confidence is low. Finally, extending the framework to spatio-temporal localization, multi-label activity recognition, and open-set detection on newer datasets (HMDB-51, UCF-101, Charades, Something-Something V2, Kinetics) will establish broader external validity and clarify how the method scales to richer taxonomies and real-world video workloads.

Data Availability Statement

The KTH, Hollywood-2, and UCF-50 datasets analyzed in this study are publicly available from their original providers. The code and trained models are available from the corresponding author upon reasonable request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Aggarwal, J. K., & Ryoo, M. S. (2011). Human activity analysis: A review. *Acm Computing Surveys (Csur)*, *43*(3), 1-43. [CrossRef]

[2] Pareek, P., & Thakkar, A. (2021). A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review, 54*(3), 2259-2322. [CrossRef]

[3] Ali, S., & Shah, M. (2008). Human action recognition in videos using kinematic features and multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence, 32*(2), 288-303. [CrossRef]

[4] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016, September). Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision* (pp. 20-36). Cham: Springer International Publishing. [CrossRef]

[5] Blank, M., Gorelick, L., Shechtman, E., Irani, M., & Basri, R. (2005, October). Actions as space-time shapes. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1* (Vol. 2, pp. 1395-1402). IEEE. [CrossRef]

[6] Ahad, M. A. R., Tan, J., Kim, H., & Ishikawa, S. (2010, June). Action recognition by employing combined directional motion history and energy images. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops* (pp. 73-78). IEEE. [CrossRef]

[7] Calderara, S., Cucchiara, R., & Prati, A. (2008, September). Action signature: A novel holistic representation for action recognition. In *2008 IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance* (pp. 121-128). IEEE. [CrossRef]

[8] Herath, S., Harandi, M., & Porikli, F. (2017). Going deeper into action recognition: A survey. *Image and vision computing, 60*, 4-21. [CrossRef]

[9] Weinland, D., Ronfard, R., & Boyer, E. (2011). A survey of vision-based methods for action representation, segmentation and recognition. *Computer vision and image understanding, 115*(2), 224-241. [CrossRef]

[10] Chaaraoui, A. A., Climent-Pérez, P., & Flórez-Revuelta, F. (2013). Silhouette-based human action recognition using sequences of key poses. *Pattern Recognition Letters, 34*(15), 1799-1807. [CrossRef]

[11] Olatunji, I. E. (2018, August). Human activity recognition for mobile robot. In *Journal of Physics: Conference Series* (Vol. 1069, No. 1, p. 012148). IOP Publishing. [CrossRef]

[12] Franco, A., Magnani, A., & Maio, D. (2020). A multimodal approach for human activity recognition based on skeleton and RGB data. *Pattern Recognition Letters, 131*, 293-299. [CrossRef]

[13] Davis, J. W. (2001, July). Hierarchical motion history images for recognizing human motion. In *Proceedings IEEE Workshop on Detection and Recognition of Events in Video* (pp. 39-46). IEEE. [CrossRef]

[14] Davis, J. W., & Bobick, A. F. (1997, June). The representation and recognition of human movement using temporal templates. In *Proceedings of IEEE computer society conference on computer vision and pattern recognition* (pp. 928-934). IEEE. [CrossRef]

[15] Papadopoulos, G. T., Axenopoulos, A., & Daras, P. (2014, January). Real-time skeleton-tracking-based human action recognition using kinect data. In *International conference on multimedia modeling* (pp. 473-483). Cham: Springer International Publishing. [CrossRef]

[16] Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D

convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(01), 221-231. [CrossRef]

[17] Dollár, P., Rabaud, V., Cottrell, G., & Belongie, S. (2005, October). Behavior recognition via sparse spatio-temporal features. In *2005 IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance* (pp. 65-72). IEEE. [CrossRef]

[18] Vemulapalli, R., Arrate, F., & Chellappa, R. (2014, June). Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group. In *2014 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 588-595). IEEE. [CrossRef]

[19] Cippitelli, E., Gasparrini, S., Gambi, E., & Spinsante, S. (2016). A human activity recognition system using skeleton data from RGBD sensors. *Computational intelligence and neuroscience, 2016*(1), 4351435. [CrossRef]

[20] Marszalek, M., Laptev, I., & Schmid, C. (2009). Actions in context. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2929-2936). IEEE. [CrossRef]

[21] He, Y., Shirakabe, S., Satoh, Y., & Kataoka, H. (2016, October). Human action recognition without human. In *European Conference on Computer Vision* (pp. 11-17). Cham: Springer International Publishing. [CrossRef]

[22] Jain, M., Jégou, H., & Bouthemy, P. (2016). Improved motion description for action classification. *Frontiers in ICT, 2*, 28. [CrossRef]

[23] Cai, J. X., Feng, G. C., & Tang, X. (2013, September). Human action recognition using oriented holistic feature. In *2013 IEEE International Conference on Image Processing* (pp. 2420-2424). IEEE. [CrossRef]

[24] Wang, H., & Schmid, C. (2013). Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision* (pp. 3551-3558). [CrossRef]

[25] Junejo, I. N., Dexter, E., Laptev, I., & Pérez, P. (2011). View-independent action recognition from temporal self-similarities. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 33*(1), 172-185. [CrossRef]

[26] Niebles, J. C., Wang, H., & Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. *International journal of computer vision, 79*(3), 299-318. [CrossRef]

[27] Ke, Y., Sukthankar, R., & Hebert, M. (2005, October). Efficient visual event detection using volumetric features. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1* (Vol. 1, pp. 166-173). IEEE. [CrossRef]

[28] Kovashka, A., & Grauman, K. (2010, June). Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 2046-2053). IEEE. [CrossRef]

[29] Shi, L., Zhang, Y., Cheng, J., & Lu, H. (2020). Skeleton-based action recognition with multi-stream adaptive graph convolutional networks. *IEEE Transactions on Image Processing, 29*, 9532-9545. [CrossRef]

[30] Poppe, R. (2010). A survey on vision-based human action recognition. *Image and vision computing, 28*(6), 976-990. [CrossRef]

[31] Xu, Z., Zhang, J., Zhang, P., & Ding, P. (2024). Skeleton-weighted and multi-scale temporal-driven network for video action recognition. *Journal of Electronic Imaging, 33*(6), 063056-063056. [CrossRef]

[32] Li, H., Huang, J., Zhou, M., Shi, Q., & Fei, Q. (2022). Self-attention pooling-based long-term temporal network for action recognition. *IEEE Transactions on Cognitive and Developmental Systems, 15*(1), 65-77. [CrossRef]

[33] Chen, J., & Ho, C. M. (2022, January). MM-ViT: Multi-Modal Video Transformer for Compressed Video Action Recognition. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 786-797). IEEE. [CrossRef]

[34] Noor, N., & Park, I. K. (2023, October). A Lightweight Skeleton-Based 3D-CNN for Real-Time Fall Detection and Action Recognition. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (pp. 2171-2180). IEEE. [CrossRef]

[35] Hong, Z., Li, Z., Zhong, S., Lyu, W., Wang, H., Ding, Y., ... & Zhang, D. (2024). Crosshar: Generalizing cross-dataset human activity recognition via hierarchical self-supervised pretraining. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 8*(2), 1-26. [CrossRef]

[36] Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., & Baik, S. W. (2017). Action recognition in video sequences using deep bi-directional LSTM with CNN features. *IEEE access, 6*, 1155-1166. [CrossRef]

[37] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems, 27*.

[38] Kilic, U., Karadag, O. O., & Ozyer, G. T. (2025). AGMS-GCN: Attention-guided multi-scale graph convolutional networks for skeleton-based action recognition. *Knowledge-Based Systems, 311*, 113045. [CrossRef]

[39] Shabaninia, E., Nezamabadi-pour, H., & Shafizadegan, F. (2024). Multimodal action recognition: a comprehensive survey on temporal modeling. *Multimedia Tools and Applications, 83*(20), 59439-59489. [CrossRef]

[40] Alsarhan, T., Ali, S. S., Ganapathi, I. I., Ali, A., & Werghi, N. (2024). PH-GCN: Boosting human action recognition through multi-level granularity with pair-wise hyper GCN. *IEEE Access*. [CrossRef]

[41] Xia, L., Chen, C. C., & Aggarwal, J. K. (2012,

June). View invariant human action recognition using histograms of 3d joints. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops* (pp. 20-27). IEEE. [CrossRef]

[42] Geng, T., Zheng, F., Hou, X., Lu, K., Qi, G. J., & Shao, L. (2022). Spatial-temporal pyramid graph reasoning for action recognition. *IEEE Transactions on Image Processing*, *31*, 5484-5497. [CrossRef]

[43] Dong, Z., Xie, M., & Li, X. (2023). Multi-scale receptive fields convolutional network for action recognition. *Applied Sciences*, *13*(6), 3403. [CrossRef]

[44] Li, J., & Yang, Z. (2024, September). Human action recognition using CNN and ViT. In *2024 International Conference on Electronics and Devices, Computational Science (ICEDCS)* (pp. 107-111). IEEE. [CrossRef]

[45] Wu, H., Ma, X., & Li, Y. (2023). Multi-level channel attention excitation network for human action recognition in videos. *Signal Processing: Image Communication*, *114*, 116940. [CrossRef]

[46] Wu, X., Zhu, J., & Yang, L. (2024). Faster-slow network fused with enhanced fine-grained features for action recognition. *Journal of Visual Communication and Image Representation*, *105*, 104328. [CrossRef]

[47] Tejonidhi, M. R., Aravinda, C. V., Kumar, S. A., Madhu, C. K., & Vinod, A. M. (2025). Optimizing group activity recognition with actor relation graphs and GCN-LSTM architectures. *IEEE Access*. [CrossRef]

[48] Liu, B., Zheng, T., Zheng, P., Liu, D., Qu, X., Gao, J., ... & Wang, X. (2023, October). Lite-mkd: A multi-modal knowledge distillation framework for lightweight few-shot action recognition. In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 7283-7294). [CrossRef]

[49] Xu, M., Xiong, Y., Chen, H., Li, X., Xia, W., Tu, Z., & Soatto, S. (2021). Long short-term transformer for online action detection. *Advances in Neural Information Processing Systems*, *34*, 1086-1099.

[50] Guo, G., & Lai, A. (2014). A survey on still image based human action recognition. *Pattern Recognition*, *47*(10), 3343-3361. [CrossRef]

[51] Reddy, K. K., & Shah, M. (2013). Recognizing 50 human action categories of web videos. *Machine vision and applications*, *24*(5), 971-981. [CrossRef]

[52] Hussain, A. (2025). Detection and Recognition of Real-Time Violence and Human Actions Recognition in Surveillance using Lightweight MobileNet Model. *ICCK Journal of Image Analysis and Processing*, *1*(3), 125-146. [CrossRef]

[53] Ji, X., & Liu, H. (2009). Advances in view-invariant human motion analysis: A review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(1), 13-24. [CrossRef]

[54] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.[CrossRef]

[55] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780. [CrossRef]

Hidayat Ullah Khan received his Bachelor & Master Degrees in Computer Science from The University of Agriculture Peshawar, Pakistan in 2017, respectively. He has more than 7 years of teaching & research experience. His Research interest includes Machine Learning, Computer Vision, Wireless Networks, Sensor Networks, Smart Healthcare, Underwater Acoustic Networks, IoT Security. (Email: hidayatkhanofficial73@gmail.com)

Altaf Hussain received his Bachelor Degree in Computer Science from University of Peshawar, Pakistan in 2013 & Master Degree in Computer Science from The University of Agriculture Peshawar, Pakistan in 2017, respectively. He has more than 6 years of teaching & research experience. He worked at The University of Agriculture Peshawar in Faculty of IT as Researcher from 2017 to 2019. He has supervised many bachelor's and master's degree level students and helped them with their final year projects and research. During his Master study, he has completed his research in drone communication systems. Currently, he is a PhD Scholar in School of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China, specializing in Underwater Acoustic Sensor Networks with focus on energy efficiency, channel optimization, localization and dynamic scheduling integrating 6G and quantum computing. He has served as a Lecturer in Computer Science Department in Government Degree College Lal Qilla Dir Lower, KPK Pakistan from 2020 to 2021. He has worked as Research Assistant with the Department of Accounting and Information Systems, College of Business and Economics, Qatar University, Doha, Qatar. He also worked as IT clerk in the Court of District and Session Judge Timergara Dir Lower from 2022 to 2023. He has published more than 30 notable journal articles. He has reviewed many articles and is serving as reviewer for Cluster Computing, Computing, Cybernetics and Systems, Journal of Cloud Computing, Knowledge and Information Systems, Peer-to-Peer Networking and Applications, SN Applied Sciences, The Imaging Science Journal, The Journal of Supercomputing, Transactions on Emerging Telecommunications Technologies, Wireless Personal Communications, Frontiers in Big Data, CMC-Computers, Materials & Continua, and Bulletin of Electrical Engineering and Informatics (BEEI). His Research interest includes Artificial Intelligence, Gesture Detection, Wireless Networks, Underwater Sensor Networks, Unmanned Drone Communication Systems, Blockchain, Federated Learning, Reinforcement Learning, Quantum Computing, IoT, and 6G Networks. (Email: altafkfm74@gmail.com)