DESIGN AND DEVELOPMENT OF PREDICTIVE MODEL FOR ACTIVITY RECOGNITION USING MACHINE LEARNING

A Thesis
Submitted in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in

Computer Science

by

ROSHNI SINGH
(2K21/PHDCS/502)

Under the supervision of
Dr. Abhilasha Sharma
Delhi Technological University

Department of Software Engineering
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi 110042

December, 2025

ACKNOWLEDGEMENTS

कर्पूरगौरं करुणावतारं, संसारसारं भुजगेन्द्रहारम् ।
सदावसन्तं हृदयारविन्दे, भवं भवानीसहितं नमामि ॥

At the core of my academic odyssey, I am deeply grateful to the Almighty for blessings, perseverance, and inspiration during the journey. In both moments of challenge and triumph, faith has provided me the fortitude to navigate the complexities of this journey. The belief in a higher purpose has remained a constant beacon, offering solace and inspiration throughout my research endeavors.

Foremost, I wish to express my profound appreciation and regards to my supervisor, **Dr. Abhilasha Sharma**, for her unwavering support, belief in my abilities, and comprehensive guidance throughout my doctoral studies. Her deep interest in the research, patient mentorship, and insightful feedback have been pivotal in shaping the quality and direction of this work. I remain truly grateful for her invaluable mentorship. I extend my heartfelt gratitude to Delhi Technological University for selecting me as a candidate for this course. I am deeply appreciative of **Prof. Prateek Sharma**, Vice-Chancellor of Delhi Technological University, Delhi, India, for his unwavering encouragement. His constant support has inspired young researchers like me to strive for excellence in academic and research pursuits. I also wish to express my profound gratitude to **Prof. Ruchika Malhotra**, Head of the Department of Software Engineering at Delhi Technological University, for her steadfast guidance and motivation throughout my research.

I extend my deepest gratitude to my parents and brothers for their unwavering love, encouragement, and understanding, which have formed the foundation of my academic and personal growth. To my mother, **Mrs Prem Geeta Singh**, thank you for your emotional strength and endless faith in me always. Your resilience has inspired me every step of the way. To my sister, Diksha Kurchaniya, your unwavering support and comforting presence have been my quiet strength. Your encouragement during difficult moments provided a sense of peace and motivation that sustained me throughout this journey.

I am also thankful to my friends, whose encouragement and companionship helped me persevere through the challenges and made the journey more fulfilling and memorable. The shared experiences and laughter added balance to my academic life. Sincere thanks are due to my colleagues, peers, and the faculty and staff at Delhi Technological University. Their support, camaraderie, and willingness to share knowledge created a nurturing environment for my research and personal development. The intellectual exchanges within this vibrant academic community were invaluable to my growth.

This thesis is the culmination of not just my work, but the collective support, love, and encouragement of all these remarkable individuals. I am profoundly grateful for their contributions, which have had a lasting impact on my academic and personal journey.

Place: Delhi Roshni Singh

Date: December 10, 2025

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahabad Daulatpur, Main Bawana Road, Delhi-110042, INDIA

CANDIDATE'S DECLARATION

I **Roshni Singh, (2K21/PHDCS/502)** hereby declare that the work which is being presented in the thesis entitled **" Design and Development of Predictive Model for Activity Recognition using Machine Learning"** in the partial fulfillment of the requirements for the award of the Degree of Doctor of Philosophy, submitted in the Department of Software Engineering, Delhi Technological University is an authentic record of my own work carried out during the period from 2022 to 2025 under the supervision of Dr. Abhilasha Sharma.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Roshni Singh
(2K21/PHDCS/502)

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of my knowledge.

Signature of Supervisor **Signature of External Examiner**

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahabad Daulatpur, Main Bawana Road, Delhi-110042, INDIA

CERTIFICATE BY THE SUPERVISOR

Certified that **Roshni Singh (2K21/PHDCS/502)** has carried out her research work presented in this thesis entitled **"Design and Development of Predictive Model for Activity Recognition using Machine Learning"** for the award of **Doctor of Philosophy** from Department of Software Engineering, Delhi Technological University, Delhi, under my supervision. The thesis embodies results of original work, and studies are carried out by the student herself and the contents of the thesis do not form the basis of the award of any other degree to the candidate or to anybody else from this or any other institution.

Dr. Abhilasha Sharma
(Supervisor)
Department of Software Engineering
Delhi Technological University, Delhi

Date: December 10, 2025
Place: New Delhi

ABSTRACT

This thesis aims to develop robust and adaptive methods for activity recognition in challenging environments involving occlusions, low visibility, complex motion patterns, and dynamic backgrounds. While conventional methods have shown promise using handcrafted features or template-based models, their performance significantly degrades under real-world conditions due to limitations in generalization, sensitivity to noise, and dependency on clean, well-labeled datasets. To address these issues, the work explores multiple directions, including skeleton-based recognition, spatial-temporal modeling, attention mechanisms, low-light enhancement, and multimodal fusion. The proposed methods are designed to enhance both the accuracy and robustness of recognition systems in real-world settings.

Initially, the thesis outlines a systematic literature review that analyzes 88 key publications from 2014 to 2024, selected from over 8,664 research papers. This review categorizes state-of-the-art HAR techniques based on their architectures, datasets, evaluation strategies, and challenges, highlighting the research gaps in handling real-time, noisy, and occluded scenarios. Based on these insights, a set of machine learning and deep learning frameworks, models and algorithms are proposed.

The second work introduced a ConvST-LSTM-Net for skeleton-based activity recognition. This model identifies and processes only the most informative skeletal keyjoints in each frame, leveraging convolutional and spatiotemporal LSTM layers for effective long-term sequence modeling. To capture subtle spatial-temporal variations in video clips, a spatial-temporal attention-based, i.e., STAD-ConvBi-LSTM model is developed in the third work. This architecture integrates a dual attention mechanism with convolutional and bi-directional LSTM networks to extract discriminative human-centric features. The method demonstrates exceptional performance across various datasets and a custom synthetic dataset, achieving recognition accuracies exceeding 96%. For recognizing the challenge of occlusion in skeleton-based data, a Multi-Stream Part-Aware Spatial-Temporal Graph Convolutional Network as MSPAST-GCN is proposed. This model uses a part-aware inhibition strategy and a graph convolution-based architecture to effectively model keyjoint relationships, even in the presence of missing or noisy data. It outperforms prior methods with a 6% accuracy gain on occlusion-affected datasets. For video-based activity classification, a hybrid model named MV-DBiLSTM is presented, which combines MobileNetV2 for spatial feature extraction with a Deep Bi-LSTM network for learning temporal dependencies. This framework balances computational efficiency and deep temporal reasoning, making it

suitable for deployment in smart systems. In visually challenging conditions like low-light environments, where traditional recognition systems face challenges. This thesis proposes a low-light enhancement pipeline integrated with HAR models. A combination of local enhancement modules and transformer-based global adjustment is used to improve visibility without distorting critical features. This significantly improves activity detection in surveillance scenarios under poor lighting.

All proposed models are rigorously validated across benchmark and synthetic datasets using both quantitative and qualitative assessments. The analysis demonstrates that all the presented methods outperform contemporary approaches in terms of recognition accuracy, temporal consistency, and adaptability to diverse real-world conditions. Overall, this thesis contributes multiple novel activity recognition architectures tailored for different challenges: occlusion, temporal complexity, lighting conditions, and data constraints. These contributions enable the development of more smart, intelligent, reliable, and context-aware recognition systems, with impactful applications in surveillance, healthcare, smart homes, and assistive technologies.

PUBLICATIONS

Published/ Accepted

SCI-INDEXED JOURNALS

1. Roshni Singh, Abhilasha Sharma."ConvST-LSTM-Net: convolutional spatiotemporal LSTM networks for skeleton-based human action recognition" **International Journal of Multimedia Information Retrieval**, 12,34 (2023). https://doi.org/10.1007/s13735-023-00301-9. (*Impact Factor-2.9).*

2. Roshni Singh, and Abhilasha Sharma."STAD-ConvBi-LSTM: Spatio-temporal attention-based deep convolutional Bi-LSTM framework for abnormal activity recognition." **Journal of Visual Communication and Image Representation** (2025): 104465. https://doi.org/10.1016/j.jvcir.2025.104465 (*Impact Factor-3.1).*

3. Roshni Singh, and Abhilasha Sharma. "Occluded skeleton-based multi-stream model using Part-Aware Spatial–Temporal Graph Convolutional Network for human activity recognition." **Engineering Applications of Artificial Intelligence** 156(2025): 111183. https://doi.org/10.1016/j.engappai.2025.111183 (*Impact Factor-8.0).*

Communicated/Under Review

SCI-INDEXED JOURNALS

1. Roshni Singh, Abhilasha Sharma, "Deep Learning for Human Activity Recognition: A Systematic Review of a Decade of Progress", in **Knowledge and Information Systems**. (Under Review) (*Impact Factor-2.5).*

2. Roshni Singh, Abhilasha Sharma, "Spatio-Temporal Siamese Ensemble with Multi-Level Feature Fusion for Video-Based Person Re-Identification", in **Expert Systems with Applications**. (Communicated) (*Impact Factor-7.5)*.

3. Roshni Singh, Abhilasha Sharma, " Entropy-Aware Cricket Activity Recognition Using CNN and DBSCAN Clustering Computational Statistics and Data Analysis", in **Journal of Science in Sport and Exercise**. (Under Revision) (*Impact Factor-1.3)*.

CONFERENCES

SCOPUS-INDEXED JOURNALS

1. Roshni Singh, Abhilasha Sharma, "MV-DBiLSTM: An Enhanced Human Activity Recognition for Smart Surveillance Systems Using a Deep BiLSTM," 2nd International Conference on Recent Advances in Engineering and Computer Applications (ICRAECA-2025). **(Scopus, Accepted and Presented 27 Jan 2025)**.

2. Roshni Singh, and Abhilasha Sharma. "Activity Recognition in Dynamic Environments Using Image Enhancement and Vision Transformers with DETR." International Conference On Innovative Computing And Communication. Singapore: Springer Nature Singapore, 2025. https://doi.org/10.1007/978-981-96-7707-8_17

TABLE OF CONTENTS

List of Figures

List of Tables

LIST OF ABBREVIATIONS

AR Activity Recognition

HAR Human Activity Recognition

AI Artificial Intelligence

ML Machine Learning

DL Deep Learning

SLR Systematic Literature Review

CNN Convolutional Neural Network

RNN Recurrent Neural Network

SVM Support Vector Machine

LSTM Long Short-Term Memory

Bi-LSTM Bidirectional Long Short-Term Memory

GRU Gated Recurrent Unit

GAN Generative Adversarial Network

DBN Deep Belief Network

HOG Histogram of Oriented Gradient

HOF Histogram of Optical Flow

GCN Graph Convolutional Network

ST-LSTM Spatio-Temporal Long Short-Term Memory

LLE Low-Light Enhancement

MV MobileNetV2

VR Virtual Reality

HCI Human-Computer Interaction

PCA Principal Component Analysis

KNN k-Nearest Neighbor

RF Random Forest

HMM Hidden Markov Model

AUC Area Under Curve

ROC Receiver Operating Characteristic

MSE Mean Squared Error

MAE Mean Absolute Error

IMU Inertial Measurement Unit

SOTA State-Of-The-Art

FCL Fully Connected Layer

STA Spatial-Temporal Attention

GMP Global Map Pooling

GAP Global Average Pooling

ViT Vision Transformer

ConvLSTM Convolutional Long-Short Term Memory

SLR Systematic Literature Review

LLIE Low-Light Image Enhancement

TGA Transformer Global Adjustment

MLP Multi-Layer Perceptron

LBP Local Binary Patterns

DBN Deep Belief Network

GAN Generative Adversarial Network

STIP Space Time Interest Point

HMM Hidden Markov Model

RQ Reserach Question

ST-GCN Spatial-Temporal Graph Convolutional

Chapter 1

INTRODUCTION

Activity Recognition (AR), a subfield of computer vision, focuses on identifying and interpreting the actions, behaviors, or intentions of an agent, whether it is an individual, group, or object, engaged in some specific goal. When the agent is human, this process is specifically termed as Human Activity Recognition (HAR), which aims to analyze, detect, and classify activities of humans through interaction with their environment or other individuals. It involves developing algorithms capable of processing multimodal inputs, i.e., RGB, depth, and infrared, to classify, detect, and understand both coarse and fine-grained human activities within a given temporal and spatial context. This chapter provides a comprehensive background on activity recognition, including fundamental concepts, system architecture, and core terminology. It also outlines prevalent challenges in sensor-based analysis, video-based analysis and highlights a wide range of real-world applications. Further, this chapter discusses the research problem statement, key contributions, scope, and motivation behind the study, its significance, and the systematic organization of this thesis.

1.1 Activity Recognition

The rapid advancement and integration of computer vision technologies into real-world applications have significantly increased the demand for accurate and efficient algorithms capable of understanding complex human visual patterns. These extend their capability by interpreting dynamic, temporal patterns associated with human activities or behavior. These systems must not only detect and classify the normal or abnormal activities but also model motion trajectories, temporal dependencies, and contextual interactions, making it a highly challenging yet essential task in computer vision. It focuses on the automatic detection and classification of human activities using visual

data, such as images and video sequences. As a fundamental task in computer vision, AR underpins a wide range of applications in the real world, including surveillance, healthcare monitoring, assistive technologies, human-computer interaction, and smart environment systems. By enabling machines to understand human behavior in context, these systems serve as a cornerstone for developing adaptive and smart systems across diverse domains. Despite significant advancements, some challenges still exist, especially in real-world, unconstrained environments. It involves recognizing a single activity within a pre-segmented video clip and assigning it to the correct class label. Recognizing complex human activities from video data is essential for developing a wide range of practical applications. In public environments, individuals engage in various routine activities such as walking, jogging, talking, placing objects, cycling, jumping, fighting, and playing [1, 2]. Precise recognition of human activities is critical for enhancing systems in areas such as public safety, behavioral analysis, coal mines, underwater activities, and intelligent monitoring. However, these systems remain a challenging task due to several inherent complexities, including the high dimensionality of video data, variability in human motion dynamics, cluttered and dynamic backgrounds, illumination changes within the environments, and frequent occurrences of partial or full occlusion. These challenges hinder the precise recognition and classification of human activities in real-world environments. Effectively addressing these issues can unlock numerous practical applications, such as advanced surveillance systems, human-robot collaboration, sports and fitness, automatic driver monitoring systems, and real-time medical monitoring. The difficulty of AR largely depends on the nature and complexity of the activity being performed. Human activities can be categorized into four hierarchical levels based on temporal duration and structural complexity: gestures, actions, interactions, and group activities. This taxonomy is often used to illustrate the increasing levels of recognition difficulty, as depicted in Figure 1.1. Each level introduces unique challenges, particularly as interactions become more dynamic and involve multiple entities or coordination among several individuals.

Gestures: These are fine-grained, atomic movements typically involving a single body part, such as hand waving, nodding, or facial expressions. They are brief and often serve as components of more complex activities, playing a significant role in non-verbal communication.

Actions: Actions consist of a sequence of gestures performed by an individual. They are more structured and recognizable, such as walking, sitting, or throwing. Actions often serve as the building blocks for more complex behaviors.

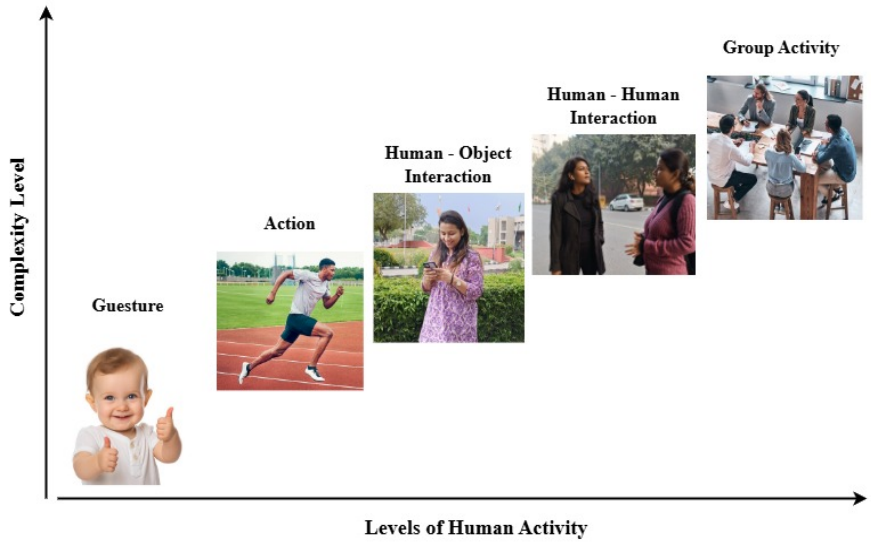


Fig. 1.1: Levels of Activity with complexity, including gesture, action, human-object interaction, human-human interaction, group activity

Interactions: These involve coordinated activities between two entities, either human-human or human-object. Examples include handshakes, hugging, using smartphones, or driving. Recognition of interactions requires understanding not only the actions of a single subject but also the spatial and temporal relationships among subjects or objects.

Group Activities: These represent the most complex category, involving coordinated activities among multiple individuals, often across dynamic environments. These activities usually involve a combination of gestures, individual actions, and interactions with people or objects. Examples team sports, protests, or group discussions.

The growing deployment of video monitoring systems in public spaces and the increasing demand for responsive, context-aware intelligent systems have made AR a central topic in vision-based Artificial Intelligence (AI) research. As such, the development of robust, real-time, and generalizable HAR models has become essential for addressing key societal challenges in various domains.

1.2 Pipeline Architecture for Activity Recognition

Activity Recognition systems typically follow a structured pipeline composed of multiple stages that process raw data into meaningful activity classifications. The efficiency and accuracy depend heavily on how well each stage in the pipeline is designed and

Fig. 1.2: Applications of Activity Recognition in Real World

optimized. The various applications of activity recognition in the real-world are shown in Fig. 1.2. Typically, AR consists of six stages, including: (1) Data acquisition for collecting data input (2) Data pre-processing, (3) Feature extraction, (4) Feature representation, (5) Activity modeling & classification, and (6) Postprocessing and Output. Each stage can be implemented using several techniques, providing the AR system with multiple choices. Consequently, selecting the appropriate application domain, data acquisition device, and machine learning models for recognition adds to the complexity of system design [3, 4]. Figure 1.3 illustrates the general workflow of an AR system. The pipeline generally includes the following key stages:

1. **Data Acquisition:** This stage involves collecting raw input data that captures human motion, which is essential for training and evaluating recognition models from various sources such as RGB cameras, depth sensors, Inertial Measurement Unit (IMU), or skeletal tracking systems [5].

2. **Preprocessing:** The collected data often contains noise or inconsistencies. Thus, pre-processing aims to enhance the quality of input video sequences for more effective feature extraction. It includes background subtraction, normalization, frame resizing, and noise filtering to enhance data quality. Techniques such as

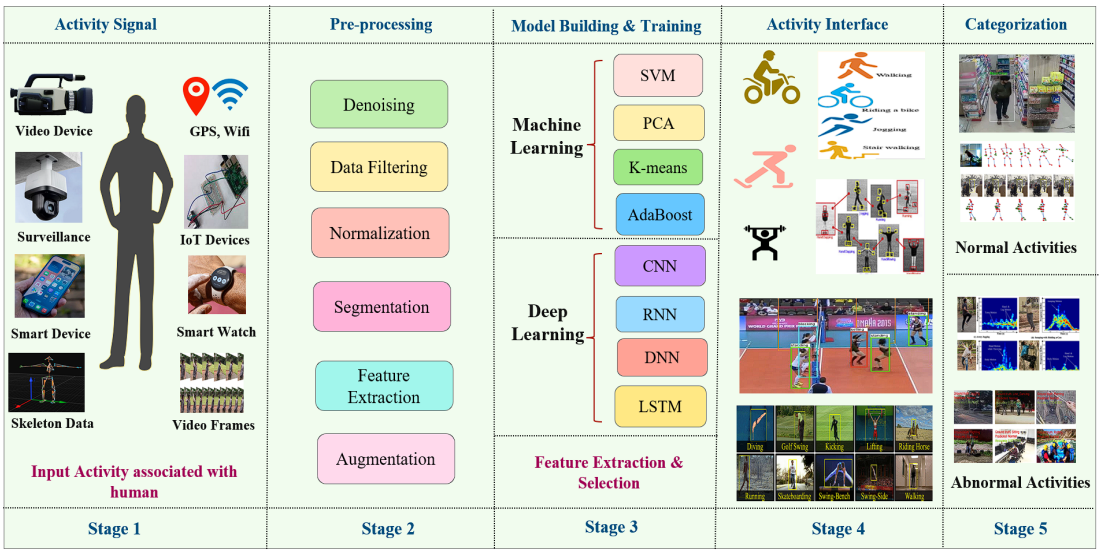


Fig. 1.3: Work Flow Diagram for Activity Recognition.

background segmentation, silhouette extraction, histogram equalization, and optical flow estimation are commonly used to improve visual clarity and motion representation [6].

3. **Feature Extraction:** This is the process of identifying meaningful features from raw data to recognize human activities effectively. At this stage, important spatial or temporal features are extracted using techniques like optical flow, keypoint detection, joint coordinates, or Convolutional Neural Network (CNN)-based feature maps [7], [8], [9], [10].

4. **Feature Representation:** Feature representation involves transforming raw input data into structured formats such as vectors, sequences, or graphs that are suitable for Machine Learning (ML) or Deep Learning (DL) models. Depending on the data type, i.e., RGB, depth, skeleton, features are extracted using handcrafted methods or learned via CNN, Recurrent Neural Network (RNN), or Graph Convolutional Network (GCN). These representations capture spatial, temporal, or relational patterns critical for accurate activity classification. The representation enhances model performance by ensuring meaningful, discriminative, and efficient input for learning.

5. **Activity Modeling and Classification:** Classification is the final and crucial step in AR pipeline, as its accuracy is highly dependent on the quality of features extracted from the input data. This stage involves using ML such as Support Vector Machine (SVM), Random Forest, or DL techniques such as Long Short-Term

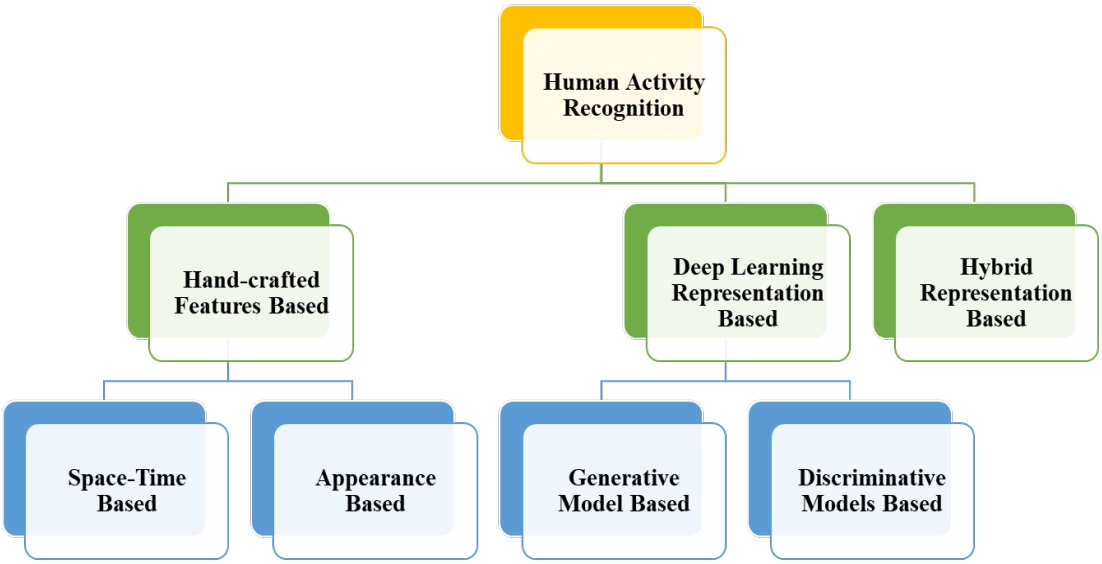


Fig. 1.4: Hierarchy of vision-based HAR approaches

Memory (LSTM), CNN, GCN, to learn patterns from extracted features and accurately recognize and categorize human activities. These models capture spatial and temporal dependencies within the data, enabling the system to distinguish between complex and similar activity classes, even under varying environmental conditions [11], [12], [13], [14], [15].

6. **Postprocessing and Output:** In some systems, postprocessing refines the predictions through smoothing or ensemble methods. The final output includes activity labels or alerts based on detected behavior.

1.3 Classification of Activity Recognition Approaches

HAR-based approaches are classified into three categories. These include (i) hand-crafted features based, (ii) DL representation-based, and (iii) hybrid representation-based approaches, as shown in Fig. 1.4.

1.3.1 Handcrafted Features Based Approaches

Handcrafted features-based action representation is the conventional way of recognizing any action. These methods analyze the video sequences or frames to extract the inclusive features and then build the descriptor. Any of the typical classifiers is then used to perform the classification task. The methods based on handmade feature design heavily rely on human inventiveness and prior knowledge. There are two kinds: space-time-based approaches and appearance-based approaches. The space-time-based approaches extract spatiotemporal features, whereas appearance-based approaches extract spatial features such as shape and motion to represent any action. The next sub-subsection discusses the various State-Of-The-Art (SOTA) approaches to HAR based on the space-time factor.

1.3.1.1 Space Time Based Approaches

Space time-based approaches characterize action representation using spatiotemporal features. They use Space Time Interest Point (STIP) detectors [7], feature descriptors, vocabulary builders, and classification modules for action representation. These detectors can be both dense and sparse. The dense STIP detector detects interest points by covering the entire video content, while the sparse detector uses subsets of the content to get interest points. The descriptors used can represent local or global features. Local Binary Patterns (LBP) [16] and E-SURF [17] use local information from an image, while Histogram of Optical Flow (HOF) [8] and Histogram of Oriented Gradient (HOG) [9] descriptors use global information for feature evaluation.

1.3.1.2 Appearance Based Approaches

Appearance-based approaches utilize visual cues derived from either shape, motion, or a combination of both to represent human activities. Shape-based methods typically rely on the contours and silhouettes of the human body to characterize actions, whereas motion-based techniques employ features like optical flow to capture movement dynamics. Some hybrid methods integrate both shape and motion features to improve recognition accuracy by leveraging complementary information. Examples include Shape Histograms [18], Body Skeleton representations [19], Motion History Volumes [20], and Optical Flow-based descriptors [21], each offering distinct perspectives on capturing and modeling human activity from visual data.

1.3.2 Deep Representation Based Approaches

Deep learning-based AR has gained significant momentum due to its outstanding performance and ability to extract meaningful features from complex, multi-dimensional data automatically. Unlike traditional ML methods, which rely heavily on handcrafted features designed manually for specific tasks, DL models are capable of learning intricate patterns directly from raw input through multiple hierarchical layers. These models require large volumes of data and substantial computational resources, particularly during training. The primary goal is to develop robust feature representations that facilitate accurate and efficient classification. Broadly, these methods can be categorized into three types: generative models, discriminative models, and hybrid approaches, each offering unique strengths in modeling human actions.

1.3.2.1 Generative Model Based Approaches

Generative models are unsupervised learning based models that represent unlabeled data distribution. Such kind of models are useful when target vectors are unavailable. It understands the data distribution with features belonging to each class and creates a new representation with compact dimensionality. Deep Belief Network (DBN) [7], Auto-encoders [22], and Generative Adversarial Network (GAN) [23] are the most commonly used generative models.

1.3.2.2 Discriminative Models-Based Approaches

Discriminatory models are supervised learning-based models that use deep hierarchies with a set of hidden layers to categorize raw inputs into related outputs. CNN or ConvNet [24] and RNN [13] are often used to perform the activity classification task.

1.3.3 Hybrid Representation Based Approaches

Some researchers show that DL architectures combined with feature-based techniques for AR. A combination of handcrafted features and deep learning approaches, including color and depth maps, as well as data, can improve the identification performance. Some approaches like HOG [9], HOF [8], Speeded-Up Robust Feature [10] etc.

1.4 Challenges in Activity Recognition

The advancements in computer vision have led to the emergence of numerous innovative recognition techniques applied to both 2D images and 3D video data. Despite this progress, accurately recognizing specific object activities remains a complex and challenging task, primarily due to factors such as variations in viewpoint, lighting conditions, partial occlusions, and intra-class diversity. While many effective approaches have been successfully adapted from image analysis to video-based AR, they often fall short in handling the complexities of real-world video content. Scenarios such as dynamic backgrounds, diverse human postures and clothing styles, and frequent occlusions continue to pose significant obstacles, highlighting the need for further improvement in recognition models for practical, real-world applications.

To address these challenges, many researchers have turned to part-based approaches, which aim to analyze only the most relevant segments of a video rather than processing it in its entirety. These segments often include trajectories, motion vectors, or spatio-temporal interest points that are indicative of meaningful activity. While part-based methods have shown promise in enhancing recognition accuracy and computational efficiency, they still face limitations. Figure 1.5 highlights some of the key issues that remain unresolved in this area. The following are some challenges in AR:

Fig. 1.5: Some existing challenges and problems faced by researchers in AR.

1.4.1 Occlusion

Occlusion occurs when parts of a subject's body are blocked either by external objects, environmental structures, or their own body movements. This issue significantly impairs the visibility of discriminative features required for accurate recognition. In dynamic scenes, such as crowded environments or cluttered indoor spaces, occlusion can cause frame-wise information loss, leading to broken pose estimations or misclassification. For effective recognition, an activity must be clearly visible throughout the video sequence. Below are the detailed sub-categories of occlusion and how they impact HAR systems:

- **Self-occlusion:** Self-occlusion occurs when different parts of a person's own body block each other from the camera's line of sight during complex actions or postures. For example: In yoga poses, the arms may fold over the torso or legs; Dancing or acrobatic movements where the limbs twist or cross frequently, etc.
- **Object occlusion:** Object occlusion occurs when external elements in the environment partially block the human subject from view. Items like chairs, desks, or doors partially obscure actions.
- **Multi-person occlusion:** This form of occlusion arises when multiple people interact or are present in a crowded scene, leading to overlapping body parts that are hard to distinguish.

1.4.2 Viewpoint Variation

Viewpoint variation refers to changes in the camera's position or angle relative to the subject performing an activity. The action view captured in the action dataset is the primary concern in identifying human action. Even minor shifts in viewpoint can drastically alter the visual representation of human motion, leading to inconsistencies in feature extraction, pose estimation, and classification. Below are the key sub-challenges that complicate AR under viewpoint variation:

- **Viewpoint Dependency:** Most AR models, especially those based on 2D vision, are highly sensitive to the angle at which an activity is observed. They often perform well on same viewpoint seen during training, but fail to generalize to unseen or novel perspectives. For example, a model trained on a front-facing view of a person walking may fail when tested on side or overhead views.

- **Lack of Multi-View Data:** Many public datasets used in this research area contain limited camera viewpoints, typically a single static angle. Multi-view or synchronized multi-camera datasets are scarce due to the complexity of data collection and synchronization.
- **Projection Distortions:** When 3D human motion is projected onto a 2D plane, as in standard RGB videos, significant spatial distortions occur. This affects the perceived scale, position, and relationship between body parts.

1.4.3 Environmental Conditions

The environmental condition can change the perspective of the scene. Due to the light sources that generate shadows on the object, a variation in illumination occurs. Weather and daytime circumstances have a tremendous impact on the scene and the artifacts formed, such as an action captured during rain, which varies radically from the identical movement taken in broad daylight or sunset. The sub-challenges are:

- **Low Illumination:** In dimly lit environments such as nighttime surveillance, indoor scenes with minimal lighting, or poorly illuminated hallways, camera sensors capture images with low contrast and higher noise. This affects the visibility of fine details are essential for activity classification.
- **Dynamic Lighting:** Environments with changing lighting conditions such as flashing lights, rotating beacons, headlights, or sudden shadows, result in frame-wise inconsistencies in pixel values. This disrupts both spatial and temporal coherence needed for motion recognition in activity sequence videos.
- **Weather Conditions:** Natural weather elements such as rain, fog, snow, or dust degrade image quality by introducing blur, occlusion, or irregular noise patterns. These artifacts mask human silhouettes and interfere with motion detection.

1.4.4 Cluttered Background

A cluttered background poses a significant challenge in vision-based action recognition, as it introduces distracting and ambiguous visual information that can interfere with the accurate interpretation of actions. In real-world applications, background scenes are rarely static or clean. They often include a wide variety of moving objects, structural patterns, or unrelated human activities. The critical sub-challenges are:

- **Visual Distraction:** Backgrounds with repetitive textures, moving elements, or other humans performing unrelated activities can visually distract models from the target subject. This reduces the clarity of motion patterns and increases the false positive rate.
- **Foreground-Background Separation:** In highly dynamic scenes, accurately separating the human subject from the background becomes difficult. The absence of static or distinguishable contrast between moving subjects and the environment impairs feature representation For example: An athlete blending into a similarly colored crowd or stadium.
- **Domain Mismatch:** Most ML-DL based HAR models are trained on curated, clean, or synthetic datasets where the background is static or simplified. When deployed in cluttered or real-world environments, these models often underperform due to domain discrepancies.

1.4.5 Inter-and Intra-Class Variability

Activities that appear visually similar can have subtle temporal differences, while the same activity performed by different individuals can look very different. This makes it hard for models to differentiate or generalize. These challenges are rooted in the diversity of human motion and the contextual similarity across different activities.

- **Intra-Class Variation:** Intra-class variation refers to the differences in how the same activity is performed by different individuals or under different conditions. These variations may include differences in body shape, speed, style, orientation, and execution dynamics, even for the same labeled action.
- **Inter-Class Similarity:** Inter-class similarity occurs when different activities exhibit overlapping motion patterns, visual appearances, or joint trajectories, leading to confusion among classes. This is especially challenging when actions share common sub-actions or involve similar limbs and motions.

1.4.6 Temporal Ambiguity and Activity Overlap

A critical challenge in HAR, especially in real-world or surveillance applications, is accurately determining when an activity starts and ends within continuous, untrimmed video streams. Temporal uncertainty can significantly degrade the model's ability to identify actions precisely and consistently.

- **Ambiguous Boundaries:** In many real-world scenarios, the exact start and end points of an activity are not clearly defined, leading to vague or overlapping transitions between consecutive actions.
- **Overlapping Actions**: In natural settings, individuals may perform multiple actions simultaneously or in quick succession, making it difficult to isolate and label each distinct activity.
- **Untrimmed Video Processing:** Untrimmed videos are long, continuous recordings where relevant activities occur sporadically. Models must automatically detect and classify actions without manual pre-segmentation.

1.4.7 Real-Time Constraints and Computational Limitations

With the growing demand for deploying HAR systems in real-time environments such as smart surveillance, autonomous vehicles, healthcare monitoring, and wearable devices, there is an urgent need to ensure that recognition models operate efficiently under stringent hardware and latency constraints. While DL has improved recognition accuracy, many of these models are computationally intensive and resource-demanding. Sub-challenges include:

- **Real-Time Processing:** Most models especially those involving CNN, RNN, or Transformers, are designed for accuracy but not necessarily for speed. In real-time HAR, the system must detect and classify activities with minimal delay or low inference time.
- **Hardware Limitations:** Unlike cloud servers with high-performance GPUs, edge devices like Raspberry Pi, NVIDIA Jetson Nano, smartphones, or embedded boards have limited memory, compute power, and thermal capacity.

1.4.8 Data Limitations and Annotation Complexity

High-quality, labeled datasets are scarce and expensive to create, particularly for complex or rare actions. Annotating long video sequences with frame-level precision is time-consuming and error-prone. These data-related challenges become especially pronounced in real-world applications, where HAR models are expected to handle variations in demographics, behaviors, attire, and environmental conditions that are often underrepresented in benchmark datasets. The sub-challenges include:

- **Lack of Diverse Datasets:** Most publicly available HAR datasets lack diversity in terms of age groups, body types, skin tones, cultural behavior, clothing styles, and interaction contexts. As a result, models trained on such datasets often fail to generalize to unseen user groups or atypical behaviors.
- **Noisy Labels:** Inconsistent, ambiguous, or erroneous annotations in training datasets often due to human labeling errors result in label noise, which hampers model training and leads to incorrect feature learning.
- **Insufficient Synthetic Data:** While synthetic data generation using tools like motion capture systems or simulation engines helps address data scarcity, there is still a significant domain gap between synthetic and real-world video data. This gap arises from discrepancies in texture, motion dynamics, background complexity, and lighting conditions.

This comprehensive set of challenges highlights the complexity and multidisciplinary nature of AR. Thus, addressing these issues requires innovative approaches that balance accuracy, efficiency, and ethical responsibility.

1.5 Motivation & Scope

With the exponential growth in visual data generated through ubiquitous camera devices, smartphones, and surveillance systems, understanding human activities from video has become increasingly vital. AR has evolved into a prominent research domain, intersecting with areas such as human motion analysis, semantic segmentation, object detection, and domain adaptation. The ability to automatically analyze and interpret human activities in video streams is crucial for a range of real-world applications. The rapid proliferation of user-generated video content on platforms like YouTube, Instagram, LinkedIn, and Twitter highlights the demand for automated, scalable, and accurate systems capable of indexing, retrieving, and understanding human behaviors in unconstrained settings. Despite the growing utility in this field, numerous challenges hinder its effectiveness. Variability in activity execution across individuals, background noise, occlusion, camera perspective shifts, and changes in lighting and appearance pose significant obstacles to accurate recognition. Furthermore, subtle activities such as gestures or complex group interactions demand context-aware and temporally consistent modeling approaches. Designing discriminative and generalizable feature representations remains a core difficulty in current AR systems.

This research is motivated by the need to develop scalable and reliable HAR models that address these challenges. By leveraging DL techniques and enhancing feature extraction mechanisms, the proposed work aims to improve the robustness and adaptability of these systems in complex real-world scenarios. The ultimate goal is to contribute toward smart systems that can perceive and respond to human behaviors in diverse settings, ranging from smart surveillance and interactive robotics to healthcare diagnostics and entertainment applications. The key motivations are as follows:

- The primary motivation behind AR lies in its wide range of real-world applications, especially in domains where human safety, security, and behavioral understanding are critical.
- Research in applications such as healthcare monitoring, intelligent surveillance, sports performance analysis, human-computer interaction, and smart environments contributes to the development of smart systems capable of understanding and responding to human behaviors in dynamic and complex scenarios.
- Offering significant societal impact by powering applications like assisted living for the elderly, early detection of abnormal or emergency behaviors, real-time threat detection in surveillance systems, and personalized, data-driven healthcare interventions.
- With the increasing availability of large-scale video data and sensor streams, there is a pressing need for scalable, accurate, and robust models capable of operating in real-time under diverse environmental conditions.
- Continued research in AR is essential to address challenges such as intra-class variability, occlusion, cluttered scenes, and temporal dependencies, enabling the deployment of generalizable models in unconstrained real-world settings.

1.6 Research Gaps

Despite the rapid advancements in AR under computer vision, current methodologies continue to face significant limitations when applied to real-world, unconstrained environments. These limitations hinder the development of systems that are both accurate and adaptive across varying conditions. A comprehensive literature review reveals the following critical research gaps:

- **Limited Viewpoint Generalization:** Most existing AR models are trained on single-viewpoint data and tend to perform poorly when activities are captured

from different or dynamic camera angles, limiting their applicability in multi-camera or real-world surveillance setups.

- **Vulnerability to Occlusion:** Current models encounter limitations in accurately recognize human activities when body parts or subjects are partially or completely occluded. This makes them less effective in crowded environments or complex scenes with object interferences.
- **Sensitivity to Illumination Variations:** Sudden changes in lighting conditions due to weather, time of day, or indoor/outdoor transitions adversely affect recognition performance. Many HAR systems lack robustness to such natural environmental fluctuations.
- **Challenges in Crowded Scenes:** Accurately identifying individual activities in crowded spaces remains a significant challenge. Existing models often misclassify or overlook actions due to overlapping subjects and cluttered backgrounds.

These gaps emphasize the need for recognition systems that are not only accurate and efficient but also robust against occlusion, lighting variance, crowd density, and viewpoint changes. They also reveal shortcomings in multi-modal data integration and generalizability across synthesized and real-world datasets highlighting areas where further research is essential.

1.7 Problem Statement

This research addresses the persistent limitations in existing activity recognition systems, including challenges such as intra-class variability, occlusion, scale and viewpoint changes, low illumination, and real-time performance constraints. These factors hinder the accuracy, robustness, and applicability of AR models in dynamic and unconstrained environments. Hence, the research's problem statement is:

"To design and develop robust and scalable methodologies for Activity Recognition that overcome the limitations of current systems, enhance classification accuracy, and enable real-time processing through advanced feature extraction, machine-learning and deep-learning-based spatio-temporal modeling."

1.8 Research Objectives

The research gaps identified in the domain of AR highlight the need for comprehensive solutions that address the complex challenges of recognition accuracy, real-time performance, robustness to occlusion and viewpoint variation, generalizability across environments, and adaptability to diverse activity types. To address these concerns, a thorough study and experimental evaluation of the available vision-based methods across multiple modalities, such as RGB, depth, and skeleton data, are crucial. Such an analysis, conducted on consistent and diverse benchmark datasets, can provide valuable insights into the strengths and limitations of existing approaches, paving the way for the development of more reliable and effective recognition models. This necessitates the exploration of innovative, lightweight architectures that can achieve a balanced trade-off between recognition capability and efficiency without compromising robustness. In response to these identified gaps, research objectives are formulated to drive progress in the field of vision-based AR. These objectives aim to foster the creation of models that are not only accurate and efficient but also scalable and adaptable to real-world conditions. The research objectives are defined as follows:

- **Objective 1**: To design a framework that can more precisely detect and classify daily living activities from existing datasets.
- **Objective 2**: To perform an extensive study to investigate the behavior of existing literature on the evaluation of synthesized datasets for activity recognition.
- **Objective 3**: To propose a multi-level feature fusion-based framework for identifying different activities in a partially occluded environment.
- **Objective 4**: To develop an adaptive and dynamic model(s) that can handle changes in environment itself.

1.9 Major Contributions of the Thesis

The major contributions include the development of a robust and scalable AR framework that effectively addresses the above-mentioned key challenges. The proposed approach integrates spatial-temporal features through DL architectures to enhance recognition accuracy and generalization.

- The core contributions of this thesis lie in the development of a novel spatio-temporal DL-architecture, ConvST-LSTM-Net, tailored for skeleton-based HAR.

This model leverages key-joint coordinates extracted from skeletal data in RGB videos, utilizing convolutional layers to capture spatial-temporal features, which are then processed through a spatio-temporal LSTM (Spatio-Temporal Long Short-Term Memory (ST-LSTM)) and time-distributed dense layers for final classification. By selectively using 17 out of 25 key joints and 21 coordinate features, the model emphasizes informative key joints, thereby enhancing recognition accuracy. The integration of CNN, Convolutional Long-Short Term Memory (ConvLSTM), and ST-LSTM paradigms forms a unified and efficient framework.

- Also, this thesis presents a novel DL-based framework for abnormal HAR using unprocessed RGB video streams as STAD-ConvBi-LSTM. The proposed model integrates a CNN for discriminative spatial feature extraction, a 6-layer bidirectional LSTM for effective long-term temporal modeling, and a dual-channel attention mechanism combining RGB and spatio-temporal cues. The channel-wise attention is strategically applied after every two convolutional layers to emphasize critical activity regions across spatial and temporal dimensions.

- This thesis presents a novel Occluded Skeleton-Based Multi-Stream Part-Aware Spatial-Temporal Graph Convolutional Network (MSPAST-GCN) for human activity recognition using skeleton input sequences. To enhance feature representation under occlusion, an inhibition strategy is introduced, enabling the model to focus on informative key points while suppressing noisy or missing joints. The core architecture, termed Part-Aware Spatial-Temporal GCN, decouples spatial and temporal modeling by applying dedicated graph convolutions to capture discriminative spatial joint correlations and dynamic temporal dependencies.

- This research introduces a Low-Light Enhancement (LLE) technique that improves illumination intensity, contrast, and color consistency in low-light environments while preserving spatial details of input images. The proposed model is rigorously evaluated on three challenging datasets SCIE, LOLO-V1, and ARID demonstrating robust prediction performance under both normal and adverse lighting conditions. Comprehensive quantitative and qualitative analyses confirm that the framework significantly outperforms existing methods in activity recognition under low-illumination scenarios.

- This research presents the MV-DBiLSTM model, which combines MobileNetV2 and Deep Bidirectional LSTM for robust human activity classification. MobileNetV2 is employed for lightweight yet effective spatial feature extraction,

while Deep Bi-LSTM captures both short- and long-term temporal dependencies by processing sequences bidirectionally. This dual-stage framework enhances recognition accuracy in complex and dynamic activity scenarios. Extensive evaluations on benchmark datasets HMDB51, UCF Sports, JHMDB, and a synthesized dataset demonstrate the model's superior performance and adaptability across diverse HAR conditions.

- Finally, this thesis concludes by outlining future research directions, including real-time processing, multimodal feature fusion, and the adoption of edge-based devices to further improve the performance. Together, the proposed methods and insights establish a solid foundation for advancing AR in complex, real-world scenarios.

1.10 Organization of Thesis

- Chapter 1: **Introduction to Activity Recognition**

  This chapter provides an overview of AR, a crucial area in computer vision that focuses on identifying and understanding human actions from visual data. With applications in surveillance, healthcare, smart homes, and human-computer interaction, and has become increasingly important in building intelligent and responsive systems. This chapter also discusses key challenges, including variations in lighting, background, occlusions, and the complexity of real-time recognition. It also traces the evolution of recognition techniques from traditional handcrafted features to advanced DL models, which offer improved accuracy and automation. The motivation and scope of the chosen research area have been discussed, followed by the formulation of a research problem statement. The unified research objectives have been identified, and the significance is further discussed.

- Chapter 2: **Literature Review**

  This chapter presents a Systematic Literature Review (SLR) of 88 selected Human Activity Recognition studies from 2014 to 2024, offering an updated overview of AR advancements using ML-DL. The studies are categorized by data modality, activity complexity, model architecture, and application area. A standard HAR system architecture is described, along with key challenges such as dataset bias, generalization issues, sensor variability, computational limits, and privacy

concerns. By critically analyzing these areas, the chapter identifies current limitations and research gaps, and the unified Reserach Question (RQ) has been divided into sub-questions leading towards certain research objectives, and the significance is further discussed. Thus, it establishes a strong foundation for the proposed work in this thesis.

- Chapters 3: **ConvST-LSTM-Net: Convolutional Spatiotemporal LSTM networks for Skeleton-based Human Action Recognition**

  This chapter presents a new class of spatio-temporal LSTM approaches named as ConvST-LSTM-Net (convolutional spatiotemporal long short-term memory network) for skeleton-based AR. The prime focus is to identify the informative key joints in each frame. The result of extensive experimental analysis exhibits that ConvST-LSTM-Net outperforms the SOTA models on various benchmark datasets for skeleton sequence data. The chapter also highlights the key contributions, including robust spatiotemporal feature extraction, end-to-end training, and the model's adaptability to challenging real-world scenarios.

- Chapter 4: **STAD-ConvBi-LSTM: Spatio-Temporal Attention-based Deep Convolutional Bi-LSTM Framework for Abnormal Activity Recognition**

  This chapter introduced an efficient novel spatial-temporal attention-based deep convolutional bi-directional long short-term memory framework (STAD-ConvBi-LSTM) that exploits human activity's prominent discriminative channel-wise spatio-temporal features. This framework also proposes a dual attentional convolutional neural network that combines a CNN model for extracting the spatial feature vector, a Bidirectional Long Short-Term Memory (Bi-LSTM) for capturing temporal dependencies, and a spatial-temporal attention mechanism for long-term modelling to extract human-centric prominent features representation in video clips. The result of extensive experimental analysis exhibits that the proposed model performs better than the SOTA in various datasets and generalizes across diverse activity scenarios.

- Chapter 5 **Occluded Skeleton-based Multi-Stream Model using Part-Aware Spatial-Temporal Graph Convolutional Network for Human Activity Recognition**

  This chapter presents a skeleton-based AR model designed to effectively handle occlusions in surveillance scenarios, where incomplete or noisy skeleton data is common. A multi-stream part-aware occluded skeleton-based GCN is designed

to improve predictions in the presence of occlusions. The model consists of three key modules: Input Inhibition Module for Skeleton Sequences, which handles incomplete or occluded skeleton data; Part-Aware Spatial-Temporal Graph Convolutional Network, which captures spatial-temporal dependencies among human body key joints; and the Predicted Score Inhibition, which refines the output by mitigating the effects of noisy data. By integrating these components, the model enhances robustness in occluded scenarios. The experiments demonstrate that the proposed method outperforms SOTA models on several benchmark datasets, achieving a 6% improvement in recognition accuracy compared to previous approaches. Additionally, the introduction of a synthesized dataset is one of the major contributions of this chapter, simulating various occlusion scenarios to enhance model training and evaluation while extracting the multi-modal features to construct more discriminative features, such as key-joint coordinates, relative coordinates, and temporal differences.

- Chapter 6 **Activity Recognition in Dynamic Environments Using Image Enhancement and Vision Transformers with DETR**

  This chapter introduces a robust and view-invariant framework for HAR in low illumination images or videos, posing challenges for human image recognition and personnel detection accuracy for detecting human abnormal activity. To address this, a Low-Light Image Enhancement (LLIE) technique for human safety and security is proposed. They utilize the local image enhancement module maps for low-light to normal light of the images at the pixel level while conserving the spatial specifics. As well, a transformer-based global adjustment module is used to refine the improved images, preventing over-brightening, under-illumination, and color distortions. Additionally, a feature similarity loss constrains target features to minimize adverse effects on detection.

- Chapter 7 **MV-DBiLSTM: An Enhanced Human Activity Recognition for Smart Surveillance Systems Using a Deep BiLSTM**

  This chapter introduces a HAR technique called MV-DBiLSTM for video datasets using a deep bi-directional LSTM combined with a CNN-based pre-trained model, MobileNetV2, for feature extraction. The process starts with using MobileNetV2 (MV) to extract deeper-level features of the video frames. After that, the features are input into an optimized DBiLSTM network to capture dependencies and process data for optimal predictions. This chapter presents an iterative procedure applied during testing that engages in fine-tuning which learned model parame-

ters, allowing us to adapt the trained model to a new environment with conditions different from those seen during training. The results demonstrate that the proposed method surpasses SOTA methods and achieves exceptional performance while delivering accurate and consistent AR in smart systems.

- Chapter 8: **Conclusion, Future Work and Social Impact:**

  This chapter provides a comprehensive summary of the proposed research, key findings, and major contributions, while also acknowledging the limitations of the study. It includes an in-depth discussion on future research directions, such as real-time deployment, edge devices for optimization, and the integration of multimodal data. Furthermore, the chapter highlights the social impact of the study, particularly its potential to enhance public safety through improved surveillance, support healthcare monitoring, and advance human-computer interaction. The chapter concludes by formally closing the thesis and reinforcing its overall significance within the field of Activity Recognition.

Chapter 2

LITERATURE REVIEW

This chapter presents an extensive SLR providing a comprehensive and up-to-date overview of HAR advancements using ML-DL techniques. The reviewed works are categorized by data modality, activity complexity, model architectures, and application domains, offering a structured understanding of activity recognition landscape. The chapter also details the architectural framework of these systems, outlining its core components and their interactions. The key challenges identified include dataset bias, limited generalizability, sensor variability, computational constraints, and privacy issues, particularly in real-world applications. Finally, this classification not only sheds light on technological progress but also reveals existing gaps in current methodologies, thereby laying the foundation for the research objectives that this thesis seeks to address.

2.1 Background

HAR systems focus on identifying, classifying, interpreting, and detecting human activities using raw data collected from various sources, such as smart devices, ambient sensors, and cameras [25, 26]. Comprehensive analysis of motion trajectories, patterns, and behavioral cues enables precise interpretation of human activities in diverse settings and environments, supporting innovation in healthcare technologies, intelligent surveillance, autonomous robotics, object tracking algorithms, and adaptive human-computer interfaces. [27, 28, 29]. Over time, it has evolved significantly, expanding its applications from healthcare to assistive technologies in smart surveillance, public safety, and crime prevention [30, 31]. In medical, motion analysis facilitates continuous patient monitoring and early detection of movement disorders like Parkinson's or Alzheimer's [32, 33, 34]. These systems enhance public safety in security and surveillance systems by recognizing suspicious behavior and potential threats in real-

time. In smart home automation by improving human-computer interaction through gesture-based or vision-based control systems. It also contributes to industrial safety by ensuring compliance with safety protocols and preventing workplace accidents through movement analysis. Technological advancements have transformed HAR systems from basic motion-sensing mechanisms to advanced DL-based frameworks capable of real-time data processing and complex pattern recognition [35, 36, 37]. The HAR can be divided into three approaches: sensor-based, vision-based, and deep learning-based. Early systems relied on wearable sensors equipped with accelerometers and gyroscopes to capture human movement [38, 39]. Later, vision-based techniques emerged for applications like pose estimation, optical flow analysis, and skeletal tracking [26, 40, 41, 42, 27, 31, 43]. More recently, DL models have applied neural networks such as CNN, RNN, LSTM networks, and Transformers to extract feature input and then classify human activities with high accuracy automatically [44, 45, 46]. Researchers in this field focus on various activities, including daily actions such as walking, brisk walking, running, jogging, jumping, sitting, and standing; sports activities like basketball, soccer, cricket, and tennis; and exercise-related movements such as gym, yoga, and aerobics. Other activities include interpersonal interactions such as high-fives, handshakes, and hugs; artistic expressions like dancing and playing musical instruments; and household tasks like dusting, chopping vegetables, cooking, and cleaning. While this is not an exhaustive list, different studies emphasize specific activity types based on their research objectives. Figure 2.1 illustrates the percentage distribution of human activities studied over the past decade. Based on the complexity of human movements, activities are generally categorized into four types: atomic-level activities, human-object interactions, interactions between individuals, and group-based activities. This review focuses on these activity types, whether performed individually or collectively, and gives an in-depth analysis of key challenges with potential improvements. Also, we discuss practical issues and provide research directions in this field.

Despite these advancements, several challenges remain, such as the inconsistency in human motion, where individual differences, environmental conditions, background cluttering, and occlusions impact model performance. Also, the lack of diverse and unbiased datasets limits the generalization capabilities of recognition models. Many existing datasets fail to represent a broad demographic spectrum. Computational complexity also presents a challenge, especially for real-time applications requiring high-speed processing on resource-limited edge devices. Furthermore, ethical concerns such as data privacy, security risks, and the potential misuse of surveillance technologies raise important questions about responsible AI deployment [47, 48, 49, 50, 51, 52]. To

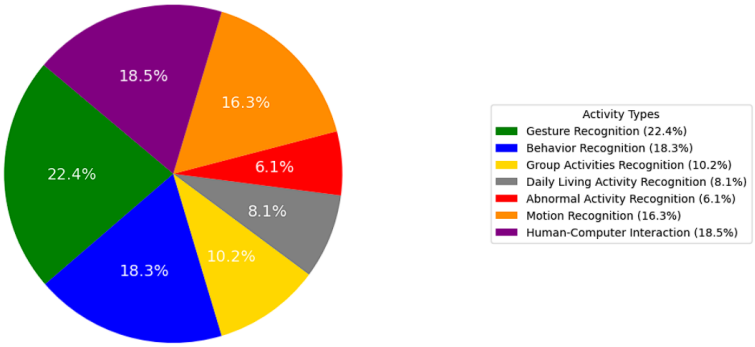


Fig. 2.1: Analysis of Studied Human Activity Types Over the Past Decade

address these issues, ongoing research focuses on improving model robustness through self-supervised learning, multi-modal data fusion, and lightweight DL architectures optimized for edge computing [53, 54, 55, 56]. The combination of emerging technologies such as Augmented Reality and Virtual Reality is further expanding the worth of recognition systems, creating more immersive and interactive applications [57, 58, 59]. Privacy-preserving AI techniques are also being developed to alleviate ethical concerns and ensure secure implementation in real-world scenarios [60, 61]. Table 2.1 shows the taxonomy of HAR categorizations based on conventional, Deep learning, unimodal data modality, and multimodal data modalities approaches, respectively. The scope of this review provides a comprehensive exploration of HAR, covering:

- The paper presents an extensive SLR of 88 selected studies from an initial pool of 8,664 articles published between 2014-2024, offering an up-to-date perspective on the evolution of HAR using Machine and Deep Learning.
- Categorization of HAR approaches based on data modality, activity complexity, ML-DL model architecture, and application domain, offering readers a structured and clear understanding of the field's landscape.
- The architectural framework of a standard HAR system, detailing key technological components and their interactions.
- The SLR highlights ongoing obstacles in research, such as dataset bias, lack of generalizability, sensor variability, computational limitations, and privacy concerns, especially in real-world deployment scenarios.
- It offers actionable research directions like domain adaptation, lightweight architectures for edge deployment, synthetic data generation, and privacy-preserving AI, encouraging innovation and guiding future studies.

Table 2.1: A Comparative Overview of Recent Surveys in HAR

Category	Ref.	Main focus
Conventional-based	[62]	Categorizing deep approaches into 2-stream, 3D-CNN, & LSTM groups.
	[63]	Reviewing extracted features in action presentation, analysis, applications, & challenges.
	[64]	Categorizing deep approaches into CNN, RNN, & hybrid groups.
	[18]	Grouping architectures into CNN, 3D CNN, RNN, & LSTM for sports video analysis.
	[65]	Grouping the methods into template, generative, & discriminative models.
	[66]	Overview of human actions feature representation using DL approaches.
	[4]	Investigating methods for abnormal action recognition.
Deep-Learning	[67]	IMU-based ZS-HAR model can recognize activities by providing skeleton videos to explain its decision-making process.
	[68]	Zero-shot video-based action recognition methods.
	[69]	combines information from two different GANs to generate visual representations of unseen classes.
	[70]	Vision Transformer and temporal modeling.
	[71]	Survey of CNNs based on input devices.
	[72]	Studying GCN.
	[73]	Two-stream architecture utilizing spatial-temporal networks to capture appearance and motion information.
	[74]	Focus on CNN-based techniques.
	[75]	HAR in video surveillance and crowd analysis.
Specific unimodal modality	[76]	semi-supervised cross-domain NN utilizing an 8×8 low-resolution infrared sensor in varying indoor environments.
	[77]	a survey of DL-methods applied to sensor-based HAR.
	[78]	Power-aware HAR using mobile and wearable sensors.

Continued on next page

Table 2.1 – continued from previous page

Category	Ref.	Main focus
	[79]	multitask DNN designed to address activity segmentation and recognition using sensor data.
	[80]	Pipeline analysis for vision- and sensor-based methods.
	[81]	Vision-based HAR by feature and input type.
	[82]	Evaluation of visual (Kinect-based) HAR.
	[83]	Deep-based motion recognition on RGB, depth, skeleton, hybrid.
Multimodal modality	[60]	Dual stream spatio-temporal using channel-wise attention network for multi-view activity recognition.
	[62]	HAR methods via RGB, wearables, and fusion approaches.
	[84]	RGB, depth, and RGB & depth modalities.
	[85]	Visual and non-visual modalities with fusion approaches.
	[86]	RGB & inertial, depth & inertial, and full fusion HAR models.
	[87]	Group activity recognition using vision, RF, and hybrids.
	[88]	Fusion strategies in C3D for RGB & depth.
	[89]	Deep and traditional methods using depth, skeleton, hybrid features.

2.2 Research Methodology

This section presents the methodological framework for conducting an SLR on HAR systems. The objective of this structured approach is to achieve thorough coverage and rigorous evaluation of relevant research within the study's scope.

2.2.1 Planning

During the initial planning phase, relevant research databases were identified, key RQ were developed, and HAR-related research articles were systematically collected. This phase established a structured approach to ensure comprehensive literature retrieval for subsequent analysis.

2.2.2 Research Questions (RQ)

The objective is to explore various human activities, actions, and patterns and then analyze their prediction. Table 2.2 summarizes the RQs formulated for this systematic survey.

2.2.3 Search Strategy

A systematic search strategy was designed to collect relevant articles comprehensively. The process involved defining appropriate search terms, establishing criteria for selecting relevant research articles and search methods, and finding credible sources to obtain. This ensured a comprehensive and unbiased selection of literature for the studies.

2.2.3.1 Proper Search Term

A systematic approach was used to identify the appropriate search terms as follows:

1. **Identification of Keywords:** The RQ were analyzed regarding popular terms, interventions, comparisons, and outcomes. The population focused on human activities, actions, motion, and movement patterns across various environments, ensuring a broad yet precise scope of research. The intervention included a variety of ML-DL methods, algorithms, data types, and datasets considered for identifying, detecting, recognizing, and predicting human activities with high accuracy. Comparisons included the year-wise review comparison based on datasets, devices used, and approaches. Finally, the outcome aimed at precisely classifying human actions using advanced DL models, data types, and standardized valuation metrics to measure the effectiveness of the recognition systems.

2. **Substitute Words, Abbreviations, and Synonyms:** Various terminologies, words, and combinations of key terms were explored:

    - *Human Activity:* "human activity recognition" OR "activity recognition" OR "human action recognition" OR "action recognition" OR "gesture recognition" OR "motion detection" OR "human pose detection" OR "pose detection" OR "daily life activities" OR "behavior analysis".
    - *Model Learning/Deep Learning* "Machine Learning" OR "ML" OR "Deep Learning" OR "DL" OR "Artificial Neural Networks" OR "ANN" OR "Convolutional Neural Networks" OR "CNN" OR "Graphical Neural Network" OR "GNN" OR "Recurrent Neural Networks" OR "RNN" OR "LSTM" OR "Transformers" OR "Attention Mechanism".

Table 2.2: Research Questions (RQs)

RQ	Research Question Statement	Motivation
RQ 1. Activities Explored		
RQ 1.1	What types of human activities are commonly detected by HAR systems?	To explore the most recent advances in HAR by identifying the types of activities that are effectively recognized using ML-DL models.
RQ 1.2	How are different categories of activities classified in HAR systems?	To assess how HAR systems categorize these activities, enhancing model effectiveness, adaptability, and real-world applicability.
RQ 2. Machine Learning and Deep Learning Methods		
RQ 2.1	Which independent & dependent variables, including input features, contribute the most to accurate activity recognition?	To determine the key variables required for effective recognition & classification of human activities.
RQ 2.2	What data representation techniques are used for pre-processing human activity data?	Understanding data preprocessing & transformation methods can enhance model accuracy and reliability.
RQ 2.3	Which ML-DL models are commonly used in HAR?	To assist researchers in selecting most efficient models for HAR applications.
RQ 2.4	What tools & open-source modules are most frequently used in HAR research?	Exploring commonly used tools can aid in improving & innovating new HAR techniques.
RQ 2.5	What training approaches are implemented in HAR models?	Optimizing training strategies ensures better generalization and robustness of recognition models.
RQ 3. Evaluation Procedures		
RQ 3.1	What benchmark datasets evaluate HAR models?	Helps in creating standardized & well-labeled datasets for HAR.
RQ 3.2	What evaluation metrics are calculated to assess model performance?	Enables comparison of various models based on standard metrics to determine their effectiveness.
RQ 3.3	How well do HAR systems generalize across different dynamic environments?	To assess the robustness and adaptability of HAR models when applied to unseen conditions or diverse users.
RQ 4. Performance Analysis		
RQ 4.1	How do dependent & independent variables have significance on activity recognition accuracy and effectiveness of feature extraction?	Understanding the impact of different factors can help improve recognition models.
RQ 4.2	What are the key factors influencing the accuracy and robustness of HAR models?	To identify the impact of data quality, feature selection, model architecture, and hyperparameter tuning on performance.
RQ 4.3	How do different ML/DL models perform in HAR across various datasets and scenarios?	Identifies the most effective techniques and provides insights for future improvements.
RQ 4.4	What is the significance of training approaches on HAR performance?	It helps to determine whether variations such as cross-dataset training yield better results.

- *Prediction:* "recognition" OR "prediction" OR "classification" OR "identification" OR "localization" OR "detection" OR "estimation" OR "model" OR "pattern recognition".
- *Data Type:* "wearable sensors" OR "IMU sensors" OR "vision data" OR "RGB cameras" OR "Skeleton-based" OR "depth sensors" OR "radar sensors" OR "multimodal data".

3. **Keyword Fine-tuning:** The pertinent keywords were carefully examined in related research papers, and those dissimilar to HAR were excluded. Keywords that did not directly contribute to human action, motion or movement, activity classification, or recognition models were removed to ensure a focused and precise search etc.

4. **Boolean Operators:** For the conjunction of related words, AND and OR were used to refine the search strategy. The OR operator combined alternative words, short forms, and synonyms of key terms. AND operator was applied to link major terms, ensuring that the retrieved literature addressed the intersection of HAR, technologies, and ML-DL methodologies.

5. **Final Search String:** The refined search string used for gathering literature was: (Human Activity Recognition OR HAR OR Action Recognition OR Motion Detection OR Behavior Analysis OR Human Pose Estimation) AND (Machine Learning OR Deep Learning OR ML OR CNN OR RNN OR GRU OR LSTM OR Transformers OR Attention Mechanism) AND (Recognition OR Identification OR Classification OR Detection OR Localization) AND (Wearable Sensors OR RGB OR Multimodal Data OR Daily life Activities OR Skeleton Data) AND (Tool OR Method OR Frameworks OR Technique OR Methodology).

2.2.4 Databases Sources

The crucial aspect of SLR is selecting the most suitable resources to identify relevant studies. To ensure comprehensive coverage of the research landscape, this review utilized well-established databases, including Springer Link, Web of Science, Scopus, IEEE Xplore, Wiley, ACM, Elsevier, arXiv, and Engineering Village.

2.3 Article Selection Process

A systematic approach was used to filter the most significant studies to ensure a focused and relevant analysis. The process used to select primary studies is illustrated in Figure 2.2 and further elaborated in the following subsections. This selection procedure ensured that only high-quality, relevant, and methodologically sound research contributions were included in the review.

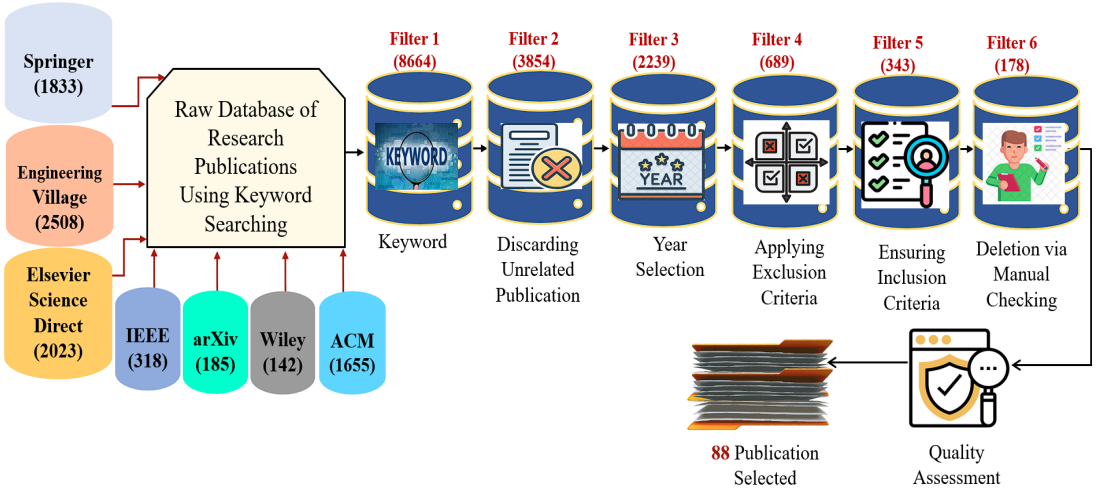


Fig. 2.2: Overview of the Article Selection Process for HAR: The step-by-step procedure followed in identifying, screening, and selecting relevant research publications for systematic literature review.

2.3.1 Databases and Search Terms Cast-off

The selection process was tangled with six filter stages to ensure that only high-quality and relevant studies were useful for defining the RQs. Prime research articles were collected using the refined search terms across multiple digital libraries, without time restrictions to ensure comprehensive coverage. The results are summarized in Table 2.3. The process of filtering is explained below:

1. In 1st filtering stage, keyword-based searches as shown in Fig. 2.3 combined with citation tracking produced an initial dataset of 8,664 journal articles.

2. In 2nd filtering stage, irrelevant papers were eliminated through assessing their relevance and research contributions.

3. The 3rd stage focused on selecting studies published after year 2014, Fig. 2.4 illustrating yearly publication trends.

4. The 4th stage applied exclusion criteria, systematically discarding papers based on title, abstract, & keywords.

5. The 5th stage applied inclusion criteria, retaining only studies aligned with the research objectives.

6. In 6th stage, the duplicate & irrelevant publications were verified and removed, with both authors collaboratively selecting 343 relevant articles approx. 0.6% of initial dataset.

7. To ensure completeness, 54 conferences & 124 journals were manually screened. Papers outside the defined scope were excluded, refining the selection to 178.

8. A final quality assessment validated the relevance and credibility of the selected papers, resulting in 88 publications meeting the criteria.

Fig. 2.3: Associated Keywords for Search

2.3.2 Inclusion and Exclusion Criteria

The inclusion & exclusion criteria were precisely defined to select studies that significantly contributed to HAR research as depicted in Table 2.4.

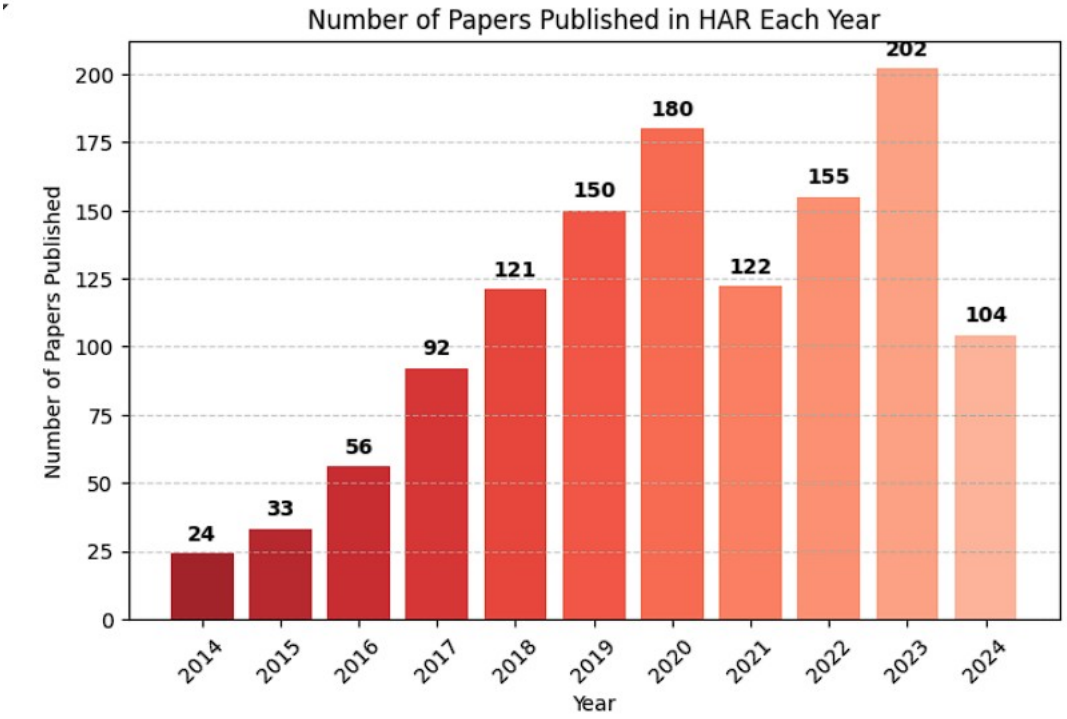


Fig. 2.4: Year-wise Publication in Last Decades (2014-2024)

2.3.3 Quality Assessment

The quality of the chosen studies was systematically assessed after the final phase. This evaluation ensured the integrity, rigor, and reliability of the selected publications. Each paper was reviewed based on predefined criteria, with responses categorized as *"Yes," "Partially," or "No"* for each question outlined in Table 2.5. A study was deemed insufficient if its methodological components could only be inferred but were not explicitly stated. To maintain accuracy, all two research authors independently evaluated each source using the quality assessment criteria. The assessment process was conducted in multiple rounds, each followed by iterative refinement. If inconsistencies emerged, the authors justified their evaluations before proceeding to the next round. This process continued until a unanimous agreement was reached, ensuring that only high-quality research studies were considered for the final analysis.

2.3.4 Data Extraction

A standardized form was used to extract data to maintain consistency and objectivity throughout the review. This form includes various details such as the authors, publication year, methodologies employed, datasets used, evaluation metrics, and principal

Table 2.3: Layer-wise Publication Filter Selection Process

Database Sources	Filter1	Filter2	Filter3	Filter4	Filter5	Filter6	Final
Springer	1833	839	457	117	75	29	20
Engineering Village	2508	1012	624	110	62	32	19
ACM	1655	724	406	211	46	18	11
Elsevier	2023	1205	548	128	78	41	15
IEEE	318	202	113	73	48	25	12
aiXver	185	85	58	31	24	27	9
Wiley	142	59	37	19	10	6	2
Total	8664	3854	2239	689	343	178	88

Table 2.4: Inclusion and Exclusion Criteria

Inclusion Criteria	Exclusion Criteria
Studies published in English from 2014 to 2024.	Non-English publications.
Research focusing on HAR within the context of Activities of Daily Living.	Duplicate studies within the databases.
Studies based on pattern data, skeleton data, or vision data.	Review articles and meta-analyses (used only for background context).
Articles published in peer-reviewed journals or conference proceedings.	Studies with inaccessible full text or published in low-quality venues.
Studies with accessible full text.	Studies with inaccessible full text or restricted access.

findings. A mixed-methods approach was then applied to synthesize the extracted data, allowing for the identification of prevailing trends, SOTA techniques, and persistent challenges. After selecting the final publications for inclusion, relevant data were extracted to address the RQs. To ensure a comprehensive understanding of past research, an additional column was included to document research gaps. This allowed new studies to identify the limitations of previous work. However, all authors collaboratively reviewed each manuscript to refine the identified limitations before finalizing the data extraction form.

2.4 Key Observations with Analysis

This section highlights the key conclusions, covering HAR system architecture, applications, datasets, techniques, and challenges identified in SLR process.

Table 2.5: Parameter Criteria for Quality Assessment Checking

Parameter Criteria	Questions	Yes	No	Partially
Objective	Is the study's objective well-defined and aligned with HAR research?	✓	X	◔
State-of-the-Art	Did the researcher evaluate their results compared to existing state-of-the-art HAR techniques?	✓	X	X
Benchmark Dataset	Are the datasets utilized in the study openly available to support reproducibility?	✓	X	◔
Accuracy	Does the proposed method effectively and accurately classify human activities?	X	X	◔
Variables	Are the dependent and independent variables for human activity recognition specified?	X	X	◔
Techniques	Is the DL/ML model or classifier definition clearly explained?	✓	X	◔
Performance Metrics	Are the evaluation metrics and performance measures explicitly disclosed?	✓	X	X
Enhancement	Are the enhancements in accuracy, robustness, or generalizability specified?	✓	X	X
Limitations	Are the limitations of the study clearly outlined and discussed?	✓	X	X

2.4.1 HAR System Architecture

HAR systems analyze raw data to interpret human actions. The process begins with collecting multimodal data from sensors [90, 64, 91] like cameras, IMUs, and smartbands, followed by pre-processing steps such as labeling, noise removal, and normalization. Features extracted from this data are then classified using ML-DL models like CNNs and RNNs [92, 44, 45, 46, 59]. This framework addresses challenges like high-dimensional data, feature consistency, and classification accuracy. Figure 2.5 gives an overview of the HAR architecture. Figure 2.6 presents a taxonomy summarizing the existing approaches within these observations. [93, 94]

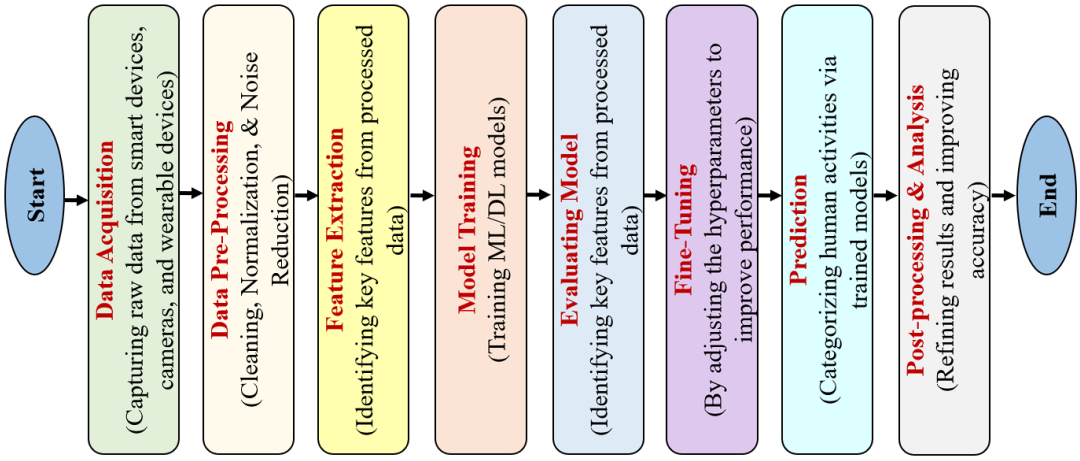


Fig. 2.5: Overview of the HAR Pipeline from Data Collection to Classification

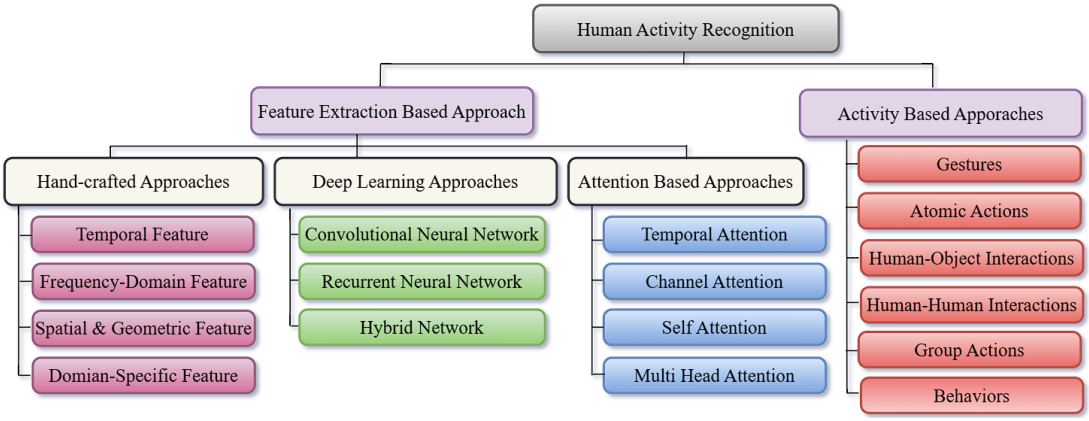


Fig. 2.6: Types of Human Activity Recognition Approaches

2.4.2 Application Areas

This section outlines the diverse areas in which these systems are applied, focusing on their significant contributions to daily living activities, surveillance, security, Human-Computer Interaction (HCI), sports, education, and robotics.

2.4.2.1 Healthcare and Daily Assisted Living

HAR is helpful in real-time applications such as monitoring patient activities, fall detection, tracking systems, tracking rehabilitation progress, and assisting the elderly healthcare centers [63, 95]. Wearable and ambient sensors are commonly used to gather movement data and identify irregular patterns that may indicate health risks. Some challenges still exist in these areas, like low-lighting environments, occlusions, camera angles, and recognizing individual activities in intricate multi-person scenar-

ios, which can affect system accuracy by up to 25%. Recent studies have shown that advanced detection and recognition models improve recognition accuracy by approximately 15%, addressing the semantic gap [90, 63, 26, 29]. The future work aims to improve data quality, model adaptability, and efficiency for seamless AR in daily life.

2.4.2.2 Human-Computer Interaction (HCI)

HAR significantly contributes to HCI by enabling gesture control in gaming, Virtual Reality (VR). It also allows for hands-free interaction with smart devices, which is particularly beneficial for disabled users [28, 96, 95]. However, real-time systems in HCI faces challenges such as fine-grained gesture recognition, latency issues, and the lack of standardized datasets for training models. Multi-modal learning, which integrates vision-based and sensor-based inputs, improves gesture recognition accuracy. Optimized DL models help in reducing processing delays. Moreover, developing benchmark datasets ensures better model generalization across diverse gestures. For future scope, integrating HAR with AR/VR technologies provides more immersive and interactive experiences. It also reduces user interaction latency while enhancing original interface.

2.4.2.3 Sports and Fitness

HAR is used to track physical performance, monitor workouts, and prevent injuries. Smart wearable gadgets like fitness bands, smartwatches, and motion sensor devices enable personalized fitness training and give motion analysis results [36, 27, 39, 97]. However, HAR models in this field often face challenges with user generalization, as individual differences in body and motion posture impact recognition accuracy. Another issue is battery life constraints in wearable devices. To overcome these issues, transfer learning allows models to adapt to different users, while pose estimation techniques enhance motion tracking. Integrating these energy-efficient AI chips in wearables also helps extend battery life without compromising performance. Challenges such as environmental factors like adverse weather conditions can reduce data reliability by up to 10%, while the complexity of athlete movements and sports equipment can further affect accuracy.

2.4.2.4 Surveillance and Security

HAR is widely used in security and surveillance to detect anomalies, analyze crowd behavior, and identify suspicious activities in public spaces [98, 99, 100]. Intelligent

surveillance systems use DL models to detect theft, violence, and unauthorized access in real-time, but raise privacy concerns due to continuous monitoring. Other challenges include low accuracy in occluded environments and high false positives. Privacy-aware approaches like skeleton or depth-based models address privacy concerns, while infrared and depth cameras improve low-light recognition. Recent advancements demonstrate a 25% improvement in detecting subtle activities, such as distinguishing routine behaviors from emergencies, a critical capability in settings like elder care, where accurate interpretation directly impacts safety [71, 27, 101]. However, challenges continue in outdoor security environments, where unpredictable weather conditions can reduce activity detection accuracy by up to 20% [99, 102, 103]. As HAR advances, its collaboration with AI will enable smarter surveillance systems, enhancing both activity detection and contextual awareness to meet real-world demands.

2.4.2.5 Smart Environments (Smart Homes, Smart Cities, IoT)

HAR is a key component in smart environments, facilitating smart home automation, smart city traffic management, and intelligent IoT applications [98, 101, 104]. In smart homes, these systems automatically adjust lighting, security, and climate control based on human presence. In smart cities, it aids in pedestrian safety and traffic monitoring. However, in smart environments must overcome challenges such as real-time adaptability, optimal sensor placement, and high energy consumption. Adaptive learning algorithms help these models continuously learn and adapt to new environments, while optimized wireless sensor networks improve data collection. The integration of energy-efficient IoT communication protocols ensures that HAR-based automation does not drain power resources excessively.

2.4.2.6 Robotics and Industrial Applications

In industrial settings, HAR enables human-robot collaboration, improving automation and safety in manufacturing processes [102, 103]. Robots equipped with HAR models recognize worker movements, assisting in tasks such as assembly and quality inspection. However, industrial HAR faces challenges related to complex movement differentiation, limited labeled datasets, and the need for high precision to ensure worker safety. GNNs are emerging as a powerful tool for recognizing structured movement data, while synthetic data generation techniques help augment datasets for training models. Explainable AI is also being explored to make these models more transparent, ensuring safe human-robot interactions in industrial environments.

2.5 Human Activity Recognition Datasets

Although the widespread use of publicly available and utilized datasets in HAR, smart living applications often demand specialized data that these datasets do not fully provide. One critical factor in choosing the correct dataset for human daily living is the relevance of activities. These datasets reflect activities pertinent to daily human routines and tasks performed in homes, workplaces, or urban settings, ensuring that adaptive intelligent systems are finely tuned to the contextual and operational needs of smart living environments. Another important factor in dataset construction is subject diversity, including participants across ages, genders, and physical abilities, which ensures a more comprehensive representation of human behavior or activities. Moreover, factors such as data quality, placement of sensor devices, and the duration of recorded activities significantly influence the performance and reliability of these models. When choosing a dataset for daily smart living applications, these aspects must be carefully evaluated. Table 2.6 shows the commonly used datasets in HAR areas with activity descriptions. Each dataset offers unique features suited to specific contexts but has limitations that require careful evaluation relative to the operational constraints and objectives of the intended application.

Table 2.6: Common Publicly Available HAR Datasets

Dataset	Samples	Activities	Description
UCI HAR [105]	Total 10,299 (7,352 train + 2,947 test)	Walking+Upstairs-Downstairs, Sitting, Standing, Laying	Accelerometer and gyroscope sensors. Environment: Indoor, supervised setting.
Kinetics-700 [106]	650,000 video-clips	700 action classes (human-object & human-human interactions)	Kinetics series (Kinetics-400, 600,700), especially with 3D CNNs & Transformers.
DAHLIA [107]	Long videos (40 min avg/subj)	Activities for smart home services (e.g., user assistance)	DAily Human Life Activity dataset recorded with KinectV2 sensors in realistic conditions with minimal instructions.

Continued on next page

Continued from previous page

Dataset	Samples	Activities	Description
Volleyball AR [108]	4830 annotated frames	Nine player actions and eight team activities	Focused on group and individual volleyball actions in sports contexts.
HHAR Dataset [109]	43,930,257 instances	Biking, Sitting, Standing, Walking, Stair-Up, Stair-Down	Data collected via smartphones & smartwatches to study sensor heterogeneities.
UTD Multimodal [110]	0.4 hours (6 participants)	Gestures and daily activities	Recorded using KinectV2 sensors for multimodal HAR tasks.
UCI-AAL [97]	5,744 sample	Sitting, Standing, Walking, Lying Down, Moving, Falling Down	Sensor data collected to improve HAR algorithms for Ambient Assisted Living applications.
SBHAR Dataset [111]	10,929 samples	Walking, Walking-Upstairs, Walking-Downstairs, Sitting, Standing, Laying	Updated version of UCI HAR dataset with inertial sensor data from smartphones.
Opportunity [112]	20 hours of recordings	Get-up, groom, relax, prepare/consume food, clean-up, open/close: doors, drawers, fridge, dishwasher; turn lights on-off, drink	Multimodal dataset for activity recognition in controlled environments using body-worn sensors.
NTU RGB+D [113]	56,880 video samples	60 action classes (e.g., daily actions, mutual actions, medical conditions)	Multimodal dataset with RGB videos, depth maps, infrared videos, and 3D skeletal data.

Continued on next page

Continued from previous page

Dataset	Samples	Activities	Description
NTU RGB+D 120 [89]	114,480 video samples	120 action classes (extension of NTU RGB+D)	Expanded version of NTU RGB+D with additional actions and samples captured by KinectV2 cameras.
PRECIS HAR [114]	800 videos	16 activities (Stand-up, Sit-down, Sit-still, Read, Write, cheer-up, Walk, Throw, Drink-bottle/mug, Move hands front/close, Raise one hand-up, Raise one leg-up, Fall-bed)	RGB-D dataset captured using the Orbbec Astra Pro camera for activity recognition in controlled environments.
CAPTURE-24 [115]	3,883 hours of accelerometer data (2,562 hours annotated)	Free-living activities (Sleep, Light-Activity, Moderate-to-Vigorous Activity, Sedentary-Behavior, Transport, Personal Care, Eating/Drinking, Social/Leisure, Unknown/Other)	Large-scale wrist-worn accelerometer dataset collected in the wild from 151 participants with wearable cameras.
KU-HAR [116]	1,945 raw samples; 20,750 sub-samples	18 activities (stand, sit, Lay, walking, running, sitting, jumping, table tennis)	Khulna University-Smartphone-based HAR dataset collected from 90 participants using accelerometer and gyroscope sensors.

Continued on next page

Continued from previous page

Dataset	Samples	Activities	Description
MPOSE2021 [100]	5,429 sequences	20 actions (e.g., walking, running, jumping)	Pose-based dataset created via OpenPose and PoseNet applied to popular RGB datasets.
ARID Dataset [117]	3,780 video clips	11 action categories (e.g., walking, drinking, waving)	AR in Dark, aims at human actions in low-light conditions to improve recognition in challenging environments.
HARTH [118]	6,461,328 instances	12 distinct such as Walking, Running, Shuffling, Stairs, Sitting, Lying	Human Activity Recognition Trondheim, Professionally-annotated dataset with accelerometer data collected in free-living settings.
SisFall [119]	4,505 samples	19: Daily Living Activity & 15: types of falls	Variety of activities, categorized into Daily Living Activities & simulated falls.
LARa [120]	3,287 annotated activity segments	Standing, Walking, Using-cart, Handling objects, Synchronization tasks	Logistic Activity Recognition Challenge taken in logistics environment from Optical marker-based Motion Capture, IMUs, RGB camera.
UT_complex [121]	1,060 labeled samples, 10 participants	3 distinct activities such as walking, jogging, biking, walking, sitting etc.	This facilitates the recognition of both simple and complex human activities using motion sensors.
UP-fall detection [65]	296,364 samples (raw sensor signals & images)	17 participants, 11 distinct activities of daily living & simulated falls	Developed to address the challenges of fair comparison between fall detection systems, it provides a rich collection of data from various sensors and devices.

Continued on next page

Continued from previous page

Dataset	Samples	Activities	Description
WISDM [122]	36 subjects (for v1.1)	Walking, Jogging, Upstairs, Downstairs, Sitting, Standing	Wireless Sensor Data Mining; Accelerometer & gyroscope readings (x, y, z axes), Sliding window techniques to create time segments.
KTH [123]	600 sequence; 25 people	Walking, Jogging, Running, Boxing, Hand Waving, Hand Clapping	Video-based HAR; Resolution 160×120 pixels, Frame Rate:25 FPS.
ActivityNet [124]	20,000 videos (849 hours total); & 200 activity categories;	Sports, Household, Hobbies, Profession and Daily activity of Pet-Dog.	Classification: Predict activity class from a trimmed clip, Temporal Localization: Predict action start/end times in untrimmed video. Detection: Multi-label action detection.
SHL [125]	4 subjects	Static: Stand-Sit-Lay; Motion: Walk-Run-Climb; Transport: Bus-Car-Subway-Train-Bike; Phone Context: In-hand, Pocket, Backpack, Not-on-body.	Human locomotion/activity recognition in real-world mobile scenarios; Real-world outdoor & indoor recordings, Real phone orientation and placement variance, Ideal for multi-device, multi-location HAR.

2.6 Result Analysis

2.6.1 Statistical Analysis

Appendix Table .1 presents the finalized selection of main publications analyzed in this SLR, including details like Year, Publication Type, journals, conferences, and citations. The data indicates that all selected publications were published between 2014

and 2024, with over 75% of these works emerging after 2020. This trend highlights the growth of DL techniques. Also, most publications were journal articles supplemented by conference papers and proceedings, indicating an increasing academic interest in this domain. A closer examination of the authorship patterns reveals that approximately 60% of these studies were co-authored, although the contributing researchers and groups varied. This analysis of key studies provides insights into the volume of research conducted, the focus areas explored, and the gaps that still require further research. Emerging researchers can leverage these findings to align their studies with the evolving needs and trends in HAR research.

2.6.2 Analysis on Research Objectives (RQ1-RQ4)

This section offers a meticulous analysis of the key attributes of publications that successfully met the inclusion, exclusion, and quality assessment criteria. It provides insight into the selection process, ensuring transparency and a solid foundation for the subsequent discussion of research trends and gaps.

RQ 1. Activities Explored: This section explores the most recent and least used advances in HAR by identifying the types of activities effectively recognized using ML-DL models.

RQ 1.1. What types of human activities are commonly detected by HAR systems? Commonly recognized activities are basic daily living activities, which are often monitored in sports, healthcare, and smart home applications to assess mobility and well-being. Healthcare applications like detecting falls, monitoring rehabilitation exercises, and identifying movement disorders. In security and surveillance, these systems recognize suspicious or anomalous behaviors such as violation, fighting, loitering, trespassing, or carrying suspicious objects, enhancing public safety through real-time alerts. In sports & fitness, we analyze athletic performance and track activities like running, jumping, or team sports actions such as passing or dribbling. Gesture-based actions like hand waving, clapping, or pointing are commonly detected to improve human-computer interaction in smart devices and VR applications. Also, these systems monitor industrial and workplace activities to ensure compliance with safety protocols by identifying improper lifting techniques or unsafe machinery operations. Furthermore, HAR systems excel in real-time monitoring scenarios to detect abnormal events like unconsciousness or crowd anomalies in public spaces. By leveraging sensor-based data and vision-based data, these systems can provide accurate AR.

RQ 1.2. How are different categories of human activities classified in HAR systems?

HAR systems classify human activities based on complexity, duration, interaction level, application context, data modality, and movement type. Activities can be classified as atomic, such as walking, sitting, jogging, or running, which involve single-step actions, or as composite, such as cooking or playing sports, which require multiple steps or interactions with objects or people. Based on duration, activities are categorized into short-term actions like clapping or waving and long-term activities like exercising or driving, which require models capable of handling sequential data. These systems also distinguish between individual activities, interactive activities involving interactions with objects or other people, and group activities. These systems are applied across various contexts, such as monitoring daily activities, healthcare tasks, security behaviors, and sports movements. Activities are typically classified by data modality, such as sensor, vision, or multi-modal. They also distinguish between static and dynamic actions. This structured classification helps select suitable datasets and algorithms for accurate recognition across diverse environments and applications.

RQ 2. Machine Learning and Deep Learning Methods: HAR research mainly uses ML-DL frameworks, focusing on variables affecting recognition, classifiers, datasets, tools, and training methods.

RQ 2.1 Which independent and dependent variables, including input features, contribute the most to accurate activity recognition?

In HAR systems, independent variables called input features & dependent variables called activity labels critically affect accuracy. Independent variables include raw sensor data and vision-based inputs. Engineered features like time and frequency-domain characteristics further boost performance. Advanced techniques such as hybrid feature extraction with CNNs, attention mechanisms, and temporal models and transformers enhance sequential learning. The quality and granularity of dependent variables also impact outcomes. Preprocessing techniques such as filtering, normalization, and sliding window segmentation that align input data with activity durations, improving reliability [25, 40, 115, 116]. Attention mechanisms further refine feature selection by suppressing irrelevant patterns [100]. Challenges such as sensor variability, environmental noise, and class imbalance can impact accuracy but can be mitigated through techniques like multi-device training, synthetic data generation, and self-supervised pre-training on diverse datasets.

RQ 2.2 What data representation techniques are used for pre-processing human activity data?

Pre-processing is a critical step as it transforms raw data into structured formats for analysis and modeling. Several data representation techniques are employed to prepare activity data for accurate prediction. Filtering methods, such as low-pass, high-pass & band-pass filters, are used to eliminate noise and isolate relevant motion signals. Feature extraction derives meaningful attributes from raw data, and spatial features such as optical flow or skeletal joint coordinates for vision-based HAR [26, 66]. Feature selection refines the dataset by removing redundant features using methods like mutual information or recursive feature elimination, improving model efficiency. Normalization techniques, such as min-max scaling or z-score, ensure uniformity across sensors and participants, while dimensionality reduction methods like Principal Component Analysis (PCA) simplify high-dimensional datasets to reduce computational costs [46]. Missing value imputation techniques address gaps in sensor data caused by device malfunctions using interpolation or k-Nearest Neighbor (KNN) methods to maintain continuity in time-series data [71, 126, 32]. Symbolic data representation techniques like symbolic aggregate approximation convert continuous time-series data into symbolic sequences for lightweight analysis on resource-constrained devices [54, 55]. Together, these pre-processing techniques ensure that human activity data is clean, structured, and optimized for ML-DL models, significantly enhancing the accuracy and reliability of these systems across diverse applications.

RQ 2.3 Which ML-DL models are commonly used in HAR?

HAR systems use various ML-DL models to classify human activities from sensor or video data. Classical ML models like Random Forest (RF), SVM, KNN, Hidden Markov Model (HMM), and Decision Trees (DT) are commonly applied to simpler tasks or smaller datasets. RF handles noisy, high-dimensional data well, SVM works for linear/non-linear tasks with careful tuning, and KNN is accurate but computationally expensive on large datasets. HMM capture sequential patterns, while DTs are interpretable but prone to overfitting in complex cases [127, 58]. DL-models, though more resource-intensive, automatically extract features, reducing the need for manual engineering. CNNs are effective for spatial feature extraction in static activities, while RNN, LSTM, and Gated Recurrent Unit (GRU) capture temporal dependencies in dynamic tasks. Hybrid models like CNN-LSTM and ConvST-LSTM combine spatial and temporal learning, achieving SOTA results [33, 102]. Transformers further enhance performance by capturing long-range dependencies. Advanced DL architectures opti-

mize spatial-temporal feature extraction while addressing challenges like sensor variability and environmental noise. Overall, classical ML remains effective for simpler HAR tasks, while DL models dominate due to their superior ability to handle complex patterns and achieve high accuracy.

RQ 2.4 What tools and open-source modules are most frequently used in HAR research?

HAR research relies on a variety of tools & open-source modules to support data pre-processing, model development, and performance evaluation. For classical ML tasks, libraries like Scikit-learn are widely used for feature engineering, model training, and evaluation, while XGBoost is preferred for its efficiency in boosting-based ensemble methods [30, 128, 120]. In DL applications, TensorFlow and PyTorch are the most popular frameworks and dynamic computation graphs, making them ideal for implementing advanced HAR models [115, 109]. OpenCV plays a critical role in vision-based tasks by providing tools for video processing, pose estimation, and integration with pre-trained models like ResNet using its DNN module [129, 130]. Some pre-trained models like ResNet, OpenPose, Mediapipe with 3D kernels, or hybrid architectures like ConvBiLSTM are frequently utilized for spatiotemporal activity recognition in both sensor and video-based HAR systems [98, 49, 44]. These tools collectively simplify the development of these systems by offering robust libraries for model implementation, efficient data processing modules.

RQ 2.5 What training approaches are implemented in HAR models?

HAR models utilize various training approaches to enhance accuracy, robustness, and adaptability across diverse datasets and activity types. Supervised learning is used for labeled datasets to train models such as CNNs, RNNs, and hybrid architectures to classify activities like walking, running, or sitting [98, 101, 130]. Hybrid models leverage the strengths of different neural networks, with CNNs extracting spatial features and LSTMs capturing temporal dependencies, while attention mechanisms further refine feature selection by focusing on relevant patterns. Reinforcement learning is a growing technology that enables models to learn optimal policies for recognition through trial-and-error interactions with the environment for dynamic and real-time tasks [129, 128]. Transfer learning is frequently implemented by pre-training models like ResNet on large, complex datasets and fine-tuning them on specific data, reducing training time and improving performance on smaller datasets [97]. Synthetic data generation using techniques like GAN addresses challenges such as class imbalance and intra-class variability by augmenting datasets and improving model generalization. Cross-validation

techniques ensure robust evaluation by testing generalizability across individuals while sliding window validation segments data into overlapping windows for better temporal representation during training. Hyperparameter tuning plays a critical role in optimizing learning rates, batch sizes, and epochs to improve model performance, with strategies like learning rate annealing ensuring smooth convergence.

RQ 3 Evaluation Procedures: Explores the evaluation methods used for HAR models, focusing on the datasets employed in this domain and the performance metrics used for result evaluation.

RQ 3.1. What benchmark datasets are cast-off to evaluate HAR models?
These models are evaluated using a variety of benchmark datasets that provide structured data for training and testing across different modalities and scenarios. Sensor-based datasets like UCI HAR [105], WISDM [122], and OPPORTUNITY [112], are widely used for wearable device applications. These datasets include accelerometer and gyroscope data capturing activities like walking, running, sitting, and daily living actions. Vision-based datasets like ARID [117], ActivityNet [124] rely on video recordings to recognize human actions. Multi-modal datasets, such as SHL [125], combine multiple sensing modalities like body-worn sensors, RGB-D cameras, and smartphones to capture richer activity representations. Specialized datasets like NTU RGB+D, NTU RGB+D 120 [113, 131] dataset focus on studying the impact of sensors on model performance in controlled environments, while ARID [117] addresses challenges in low-light conditions. These benchmark datasets help evaluate HAR models and improve their real-world performance.

RQ 3.2. What evaluation metrics are calculated to assess model performance?
HAR systems are evaluated using performance metrics like accuracy, precision, recall, F1-score, and Area Under Curve (AUC) and Receiver Operating Characteristic (ROC) to ensure robust and generalizable performance. Accuracy shows overall correctness but may be misleading on imbalanced data. Precision reduces false alarms, while recall ensures all positive instances are detected for crucial tasks like sudden fall. The F1-score balances precision and recall for better handling of imbalance. AUC-ROC helps assess binary classifications like anomaly detection. Validation methods like k-fold cross-validation and sliding window techniques further enhance reliability, especially for sensor-based, time-series data. For continuous activity prediction tasks, regression metrics like Mean Squared Error (MSE) and Mean Absolute Error (MAE) are used to measure prediction accuracy. Also, domain-specific considerations such as ad-

dressing class imbalance through under-sampling or synthetic data generation improve F1-scores and enhance model reliability. Advanced techniques like feature importance analysis using tools like SHAP provide insights into biases and feature selection strategies. Together, these metrics and methodologies ensure a comprehensive evaluation of HAR systems.

RQ 3.3. How well do HAR systems generalize across different environments and user adaptability?

HAR systems exhibit limited generalization across environments and user populations due to variations in sensor placement, environmental conditions, and inter-user movement patterns. Sensor-based models are sensitive to device orientation, with positional shifts reducing accuracy by up to 30% [118, 120, 132], while vision-based models degrade under lighting variations and occlusions. User diversity, including demographic and behavioral differences, further impacts model reliability. Cross-dataset generalization is constrained by inconsistent data formats, sampling rates, and activity taxonomies. While basic activities generalize reasonably well, complex activities often suffer from high intra-class variability. Techniques such as domain adaptation, transfer learning, and synthetic data augmentation mitigate these issues. Pre-training on large-scale datasets [89, 113] followed by domain-specific fine-tuning improves cross-domain performance. Hybrid architectures combined with attention mechanisms capture spatial-temporal dependencies for improved robustness. Despite these advances, overfitting to lab-controlled conditions remains a challenge, often inflating performance estimates. Current efforts focus on dataset standardization, lightweight model design for edge devices, and rigorous cross-domain validation to enhance real-world adaptability. Robust generalization remains a key open challenge for practical HAR deployment.

RQ 4. Performance Analysis in Depth: This provides insight into which approaches deliver SOTA results, under what conditions, and what trade-offs exist between accuracy, computational cost, and real-world applicability.

RQ 4.1 How does feature mapping representation impact activity recognition accuracy and represent the feature extracted?

Feature mapping representation crucial step in determining the accuracy by influencing how spatial-temporal patterns are extracted and utilized. Manual feature engineering relies on domain expertise to select attributes that work well for simple activities but faces difficulties with complex actions due to its limited ability to capture intri-

cate dependencies. Automated feature extraction significantly improves accuracy by autonomously learning hierarchical features from raw data. Multi-scale feature extraction methods enhance performance by employing local-global patterns, reducing confusion between similar activities like "walking upstairs or downstairs". Hybrid models combine hand-crafted features with DL representations, leveraging both domain knowledge and learned patterns to improve recognition accuracy by 3–5% over standalone models. Feature selection techniques like Joint Mutual Information Maximization optimize dimensionality by identifying the most discriminative features, reducing computational costs while maintaining high accuracy [61, 33]. Advanced methods like depthwise separable convolutions reduce computational overhead for edge deployment, while spatial features, temporal patterns, and fused spatiotemporal embeddings enable robust recognition of complex activities. Overall, automated and hybrid feature mapping approaches outperform manual methods by capturing richer patterns and improving generalization across diverse activities.

RQ 4.2 What are the key factors influencing the accuracy and robustness of HAR models?

The accuracy and robustness of HAR models are influenced by multiple factors, including validation methodologies, feature representation, model architecture, data preprocessing, sensor variability, dataset quality, environmental conditions, computational efficiency, and real-world variability. Traditional models are effective for simple actions [29, 95], while DL architectures capture spatial-temporal dependencies for complex activities, achieving accuracies above 95% [98, 99, 101, 30]. Hybrid models integrating CNNs, LSTMs, or Transformers further enhance recognition, while attention mechanisms improve robustness by prioritizing salient features [133, 115, 102]. Pre-processing steps such as filtering, normalization, and segmentation reduce noise and align temporal boundaries. Sensor variability, particularly orientation shifts, can degrade F1-scores by approximately 20% [107, 109, 103], but multi-modal fusion of IMU, GPS, and audio sensors mitigates these effects [112, 118, 120]. Dataset bias, common in lab-constrained datasets, limits generalizability, which can be addressed using synthetic data generation techniques. Environmental challenges like illumination variation and occlusion primarily affect vision-based systems. Computational efficiency is critical for edge deployment; lightweight architectures such as separable convolutions reduce parameter counts by up to 40% without compromising accuracy [115, 119, 129]. Finally, demographic diversity and device heterogeneity introduce domain shifts, addressed through domain adaptation, data augmentation, and robust validation strategies to improve real-world generalization.

RQ 4.3 How do different ML-DL models perform in HAR across various datasets and scenarios?

HAR models demonstrate varying performance across datasets and scenarios, with DL models generally outperforming traditional ML methods, especially for complex activities and large datasets. CNNs and hybrid architectures work in spatial-temporal feature extraction, achieving SOTA accuracy of 95–98% on benchmark datasets [44, 59]. Hybrid models enhance accuracy by capturing spatial-temporal dependencies, while attention mechanisms improve robustness to noise and sensor variability [130, 100]. Recurrent models are highly effective for sequential data, particularly for transitions between activities, but they require significant computational resources and large labeled datasets. In contrast, traditional ML models like RF or SVM perform well for simpler activities & achieving 85–88% accuracy on datasets [123, 89, 118] but face issues with complex actions due to their reliance on manual feature engineering. Real-world datasets present additional challenges due to sensor variability and environmental noise, leading to performance drops of 15–20% compared to lab-collected datasets [116, 122, 124]. Multi-modal datasets benefit from hybrid models that fuse data from multiple sensors, improving robustness in diverse scenarios [65]. In healthcare applications, CNNs and LSTMs excel at detecting falls or gait abnormalities with high accuracy of 97.6% [130], while synthetic data generation addresses class imbalance problem. Lightweight models or those using separable convolutions reduce computational costs by 40% [61, 99], making them suitable for edge devices. However, challenges such as subject variability (e.g., young adults vs. elderly users) and environmental factors remain significant barriers to generalization. While DL excels in controlled settings, traditional ML suits simpler for resource-limited tasks. The future work should prioritize domain adaptation, synthetic data, attention mechanisms, and robust validation for real-world deployment.

RQ 4.4 What is the significance of training approaches on HAR performance?

Training strategies directly influence HAR model generalization, robustness, and accuracy across controlled and real-world environments. Hybrid architectures such as CNN-LSTM [61] leverage spatial-temporal dependencies, outperforming homogeneous models, especially in scenarios with sensor variability [132]. Variations in device orientation and hardware heterogeneity can degrade performance by up to 45% [86], necessitating training on heterogeneous datasets. Synthetic data generation mitigates class imbalance and intra-class variability, particularly for rare activities such as falls [90]. RL introduces adaptability in dynamic environments, while hyperparameter optimization ensures optimal convergence. Advanced validation protocols expose general-

ization gaps between lab-based and real-world deployments. Sequential segmentation techniques, such as sliding windows, enhance temporal modeling [34]. Despite DL models achieving accuracies of 95–98% [126, 31], real-world factors still challenge robustness. Metrics like F1-score and maximum mean discrepancy (MMD) quantify generalization under these variations. The future directions should emphasize lightweight models for edge deployment, explainable AI for interpretability, and pre-training on diverse datasets for improved transferability. Effective training pipelines integrating hybrid architectures, data augmentation, adaptive learning, and rigorous validation are essential for bridging the gap between controlled benchmarks and real-world HAR performance.

2.7 Discussion

Machine learning and Deep learning for HAR is an emerging and continuously evolving research domain. Many studies have leveraged DL-based techniques to enhance recognition, and they have also identified significant research gaps that require further research. Several works have critically analyzed these limitations, emphasizing the need for refinement and advancement in existing methodologies. Future research directions have been the central theme, with some studies proposing specific improvements, while others have outlined novel approaches to push the boundaries of HAR innovation. This section provides a comprehensive discussion of the identified research gaps, highlighting areas that demand further exploration to advance the field.

2.8 Summary

This work presents SLR on ML-DL-based techniques for human activity recognition. The studies included in this were published between 2014 and 2024. Out of an initial pool of 8,664 articles, 88 papers were carefully selected through a rigorous selection process. These key studies were extensively analyzed, focusing on the following critical aspects:

1. Prevalent Activities in HAR: The analysis highlights commonly recognized activities in HAR datasets, such as walking, running, sitting, standing, and complex actions like hand gestures or fall detection. Studies suggest that recognizing activities in clusters enhances accuracy by leveraging shared feature representations.

2. Machine Learning and Deep Learning Approaches: The reviewed studies predominantly employ DL-based techniques such as CNNs, RNNs, LSTM, DeepLSTM, bi-LSTM, and ConvoLSTM networks. Hybrid models like CNN+LSTM and bi-LSTM demonstrated superior performance. Also, research categorized detection tools based on relevance and effectiveness. The evaluation of training methodologies showed that intra-subject training yielded better results than cross-subject training. Furthermore, feature selection techniques were analyzed to identify key metrics for HAR.

3. Evaluation Methods Applied: HAR models were assessed using publicly available datasets. The review found that accuracy, precision, recall, and F1-score were the most commonly used evaluation metrics for model performance assessment.

4. Performance Analysis: The study examined the impact of independent variables on activity recognition, revealing that fewer but well-selected features improved model performance. Comparative analysis of different DL-based architectures showed that hybrid models outperformed standalone approaches. Further, findings indicate that model efficiency is not necessarily dependent on the number of network layers, but rather on the quality of feature extraction and training strategies

This systematic review provides guidelines and recommendations for future research in HAR, along with a summary of essential feature representations for activity recognition. These insights serve as a reference for developing more robust models. The findings from this SLR suggest that this field in computer vision remains an evolving research area, with potential advancements possible through the integration of quantum computing and evolutionary algorithms for enhanced recognition accuracy.

Chapter 3

ConvST-LSTM-Net: Convolutional Spatio-Temporal LSTM Networks for Skeleton based Human Action Recognition

This chapter presents ConvST-LSTM-Net, a deep learning framework for skeleton-based human action recognition. By combining CNN and spatiotemporal LSTM units, the model aims on extracting features from only the informative keyjoints. It processes skeletal data through keypoint detection, geometric and kinematic feature extraction, and sequential modeling for action classification.

3.1 Introduction

Nowadays, crowded places with normal and abnormal activities are familiar due to population increase, which leads to suspicious activities. By recognizing and analysing the human actions from the videos, one can clearly distinguish between normal and abnormal behaviours that can make significant improvements in public safety. Withal, HAR remains an ambitious challenge due to its cluttered backgrounds, slight inter-class segregation, and wide intraclass deviation. Some applications, such as monitoring suspicious detection and early reporting for fall detection, are also considered in HAR. However, various techniques are there for the representation of human action based on motion, such as RGB-based videos, RGBD-based videos, and skeleton-based videos [134, 135]. On comparing these techniques, all the skeleton-based methods represent: (i) human motion via 3D coordinates positions for key-body points and (ii) are more robust to problems like variations of background clutter, observation viewpoints, illumination or intensity conditions and so on. These advantages of skeleton motion sequences motivate researchers to develop new techniques for exploring informative features for human action recognition. These methods are gaining utmost importance

in HAR since skeletons represent a compact sequence of data forms that depict dynamic motion within human body movements [136].

Some approaches are used for tracking and calculating the skeleton keyjoints with feature invariant to human key-body points, observation point, camera viewpoint and so on. The skeletal features within the human body are responsible for recognizing all the normal & abnormal activities. Besides this, they have also been used for evaluating (some activities such as falling, discriminating between jogging and running) the variation in the keyjoint coordinates between the center-mass of the body, acceleration-motion, velocity-motion, for movement: angles between the key-joint points within the skeleton. Some new methods like ST-LSTM and ST-GCN are practices to extract these features also. In this work, the prime objective is to efficiently combine the important cues in CNN and LSTM using Spatio-Temporal data with skeleton-based recognition approaches called ST-LSTM. Here, a set of extracted skeleton features in conjunction with skeletal keyjoint is fed as input to the model. The skeletal tracking algorithms were used for detecting the keyjoints followed by the feature extraction that was done through RGB frame data to improve the efficiency of the model. Some standard features, like angle between the keyjoint coordinates, velocity-motion, acceleration-motion, and human body position of the center of mass, for movement: angles between the joint key points, have been extracted from keyjoint coordinates of the human skeleton. Once the feature extraction is done along with pre-processing thereupon, the preprocessed data is fed to the model, consisting of 17 extracted features among 25 skeleton coordinates.

The overall pipeline of proposed ConvST-LSTM-Net model is illustrated in Fig. 3.1. The model exploits a spatio-temporal network consisting of CNNs, ST-LSTMs & fully connected dense layers. The model first detects the skeleton keyjoints of the persons using the skeleton-based recognition method. These key joints are fed to the CNN layers, followed by ST-LSTMs for the extraction of spatial-temporal features. Then, output from a hidden layer of ST-LSTMs is passed via FC dense layer for classification. The key contributions of the research work can be summarized as follows:

- A spatio-temporal ConvST-LSTM-Net model has been proposed that utilizes human body key-joint coordinates from skeletal data obtained from RGB videos. The keyjoints are fed as an input to CNN layers for extracting the spatial-temporal features, followed by ST-LSTM, and the output is passed to the time-distributed FC dense layer.

- Motivated by the advances in CNN, ConvLSTM, and ST-LSTM, we have seamlessly combined the ideas of these models and integrated them to propose a new

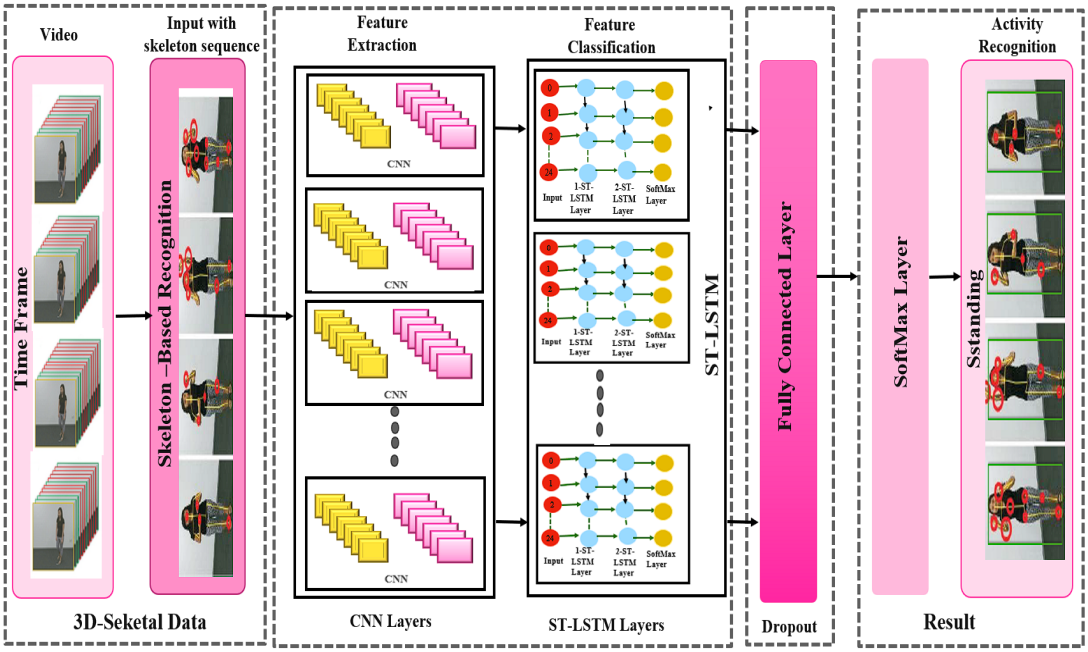


Fig. 3.1: Process Informatics Pipeline of ConvST-LSTM-Net. At first, the video frames are passed through the skeleton-based recognition feature method to extract the skeleton key joint coordinates. Then, the obtained keyjoint coordinates are fed to a modified ST-LSTM cell followed by ST-LSTM layers to evaluate the spatio-temporal feature. Further, for classification, the outputs are passed to FC-dense layers. Ultimately, SoftMax shows the framewise prediction scores of human action behaviours.

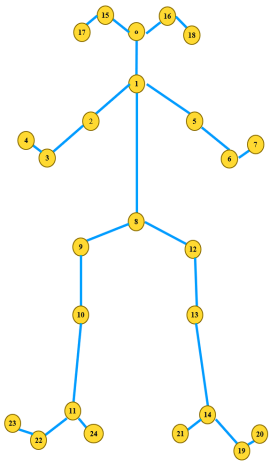
paradigm for skeleton-based action recognition termed as ConvST-LSTM-Net. The model brings attention to improving the efficacy by focusing only on informative keyjoint coordinates.

- Among 25 keyjoints, a set of 17 extracted skeleton features along with 21 skeleton keyjoint coordinates are fed to the model as not all the skeleton keyjoints are informative in nature for recognizing the action classes.
- The proposed ConvST-LSTM-Net model shows better performance in comparison to existing models by using different modalities over various benchmarks.

3.2 ConvST-LSTM-Net: The Proposed Methodology

This section briefly canvasses crucial terms and approaches used in the proposed model, which is divided into three models, namely CNN, ST-LSTM, and ConvST-LSTM-Net. The proposed methodology has been executed with the review of CNN along with the construction of skeleton-body with coordinates and ST-LSTM, respectively. Initially,

for dataset preparation, the human body frames from raw RGB videos are used to train the network while tracking the 3D skeleton joint key coordinates. For pre-processing, the 3D joints-normalization method is applied, which is helpful in making a bounding box above the tracked human body. Some features have been extracted for determining different activities, features such as velocity-motion (ν), acceleration-motion (α), weight (w), depth (d), height (H), angle (θ) within the consecutive skeleton joints, etc. After training, the feature extraction is done for input human activity behaviour.



S.No.	Key-Joint Coordinates
0	Nose
1	Neck
2	Right Shoulder
3	Right Elbow
4	Right Wrist
5	Left Shoulder
6	Left Elbow
7	Left Wrist
8	Mid Hip
9	Right Hip
10	Right Knee
11	Right Ankle
12	Left Hip
13	Left Knee
14	Left Ankle
15	Right Eye
16	Left Eye
17	Right Ear
18	Left Ear
19	Left Thumb-Toe
20	Left Little-Toe
21	Left Heel
22	Right Thumb-toe
23	Right Little-Toe
24	Right Heel

Fig. 3.2: The 25-skeleton keyjoints for the human body track detection and pre-processing.

3.2.1 Keypoint Detection & Pre-processing

In pre-processing technique for RGB videos, the frames are inputted into ST-LSTM network to evaluate the key-joint locations of human body frames from skeleton coordinates. Fig. 3.2 represents the 25-skeleton keyjoint coordinates which have been tracked at each joint. Only 17 skeletal keyjoints are covered (since they are the informative skeletal keyjoints to specify the normal and abnormal human activities) and these are the right-knee, right-hip, left-knee, left-hip, left-foot, left-ankle, right-foot, right-ankle, head, spine-mid, left-wrist, spine-base, right-shoulder, left-shoulder, right-wrist, right-elbow and left-elbow. Each frame tracked the human skeleton comprising x, y, z coordinates of human body keyjoints. After getting 3D skeletal coordinates, normalization technique has been applied on 3D keyjoints to generate bounding boxes over tracked human skeleton, which may vary as per the person's movement in video.

Afterwards, a feedforward network has been used based on a multi-CNN layer followed by ST-LSTM that takes input in the form of keyjoints coordinates from video

frames using skeleton-based recognition. It learns the affiliation among the body parts of individuals within the frames. Table 3.1 presents the details of tracked skeletal keyjoints, a set of derived features, and action class. For normalizing the convergence of loss function, minimum-maximum normalization technique (min-max norm) has been used. Here, X indicates the training dataset, then normalization can be achieved as:

$$\boldsymbol{X}_{noramlize} = \frac{\boldsymbol{X} - \boldsymbol{X}_{min}}{\boldsymbol{X}_{max} - \boldsymbol{X}_{min}} \tag{3.1}$$

Table 3.1: Detail of selected relevant skeleton keyjoints coordinates and derived features.

Lable	Description
Skeleton Joints	Right-Hip, Left-Hip, Left-Foot, Right-Foot, Right-Knee, Left-Knee, Right-Ankle, Left-Ankle, Head, Right-Wrist, Left-Wrist, Right-Shoulder, Left-Shoulder, Left-Elbow, Right-Elbow,Spine-mid, Spine-Base.
Features	$\angle\theta$, ν, $\mathfrak{a}$, hd, d, w, H, (Geometric & Kinematic features)
Action Class	Sit, Stand, Walk, Run

3.2.2 Construction & Evaluation of Feature Vector: Geometric & Kinematic Features

The skeleton keyjoint coordinates are used for constructing and calculating features vectors. The keyjoints coordinates of the human body that are tracked for different activities of humans are actually decided by using feature vectors. For particular activities different features are utilized. These features and their evaluation are as follows:

$\angle\theta$ (Angle between key-joints of skeleton coordinates): Among 25 keyjoints coordinates, we consider those coordinates which are connected via straight line and then a skeleton structure of tracked human body is drawn using these coordinates as shown in Fig. 3.3 Accordingly, only 10 keyjoint comes out to be the most relevant ones viz. left-shoulder, spine-mid, right-shoulder, spine-base, left-knee, right-knee, left-hip, right-hip, left-ankle, and right-ankle are used for calculating the value of angle θ. It is the illustration for evaluating the keyjoint angles between the left-shoulder, neck and mid-hip. If A, B, C are considered as the distance between the coordinate, then

values are formulated as $A_1 = x_1 - y_1, B_1 = x_2 - y_2$, and $C_1 = x_3 - y_3$, then θ can be evaluated as follow:

$$\boldsymbol{\theta} = \frac{\boldsymbol{ABC}}{\boldsymbol{AB} * \boldsymbol{BC}} \tag{3.2}$$

where, $ABC = (A_1 * A_2 + B_1 * B_2 + C_1 * C_2)$

$$\mathbf{AB} = \sqrt{\mathbf{A_1^2} + \mathbf{B_1^2} + \mathbf{C_1^2}} \quad \text{and} \quad \mathbf{BC} = \sqrt{\mathbf{A_2^2} + \mathbf{B_2^2} + \mathbf{C_2^2}} \tag{3.3}$$

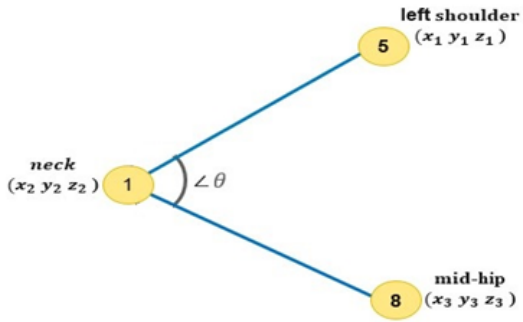


Fig. 3.3: Illustration for calculation of angle from left side of skeleton between left shoulder, neck, and mid-hip.

Velocity-Motion Estimation $(\boldsymbol{\nu})$: Velocity-motion is calculated by taking the distance of positions of humans at time frames t and $t+1$ in x, y, z -dimension. So, velocity of the tracked person ν is given in which d indicates distance of tracked person between frames and t indicates frame-time.

$$\boldsymbol{\nu} = \frac{\mathbf{d}}{\boldsymbol{t}} \tag{3.4}$$

Acceleration-Motion estimation **(a)*:*** The rate of changes of velocity between consecutive frames in x, y, z -directions at time frame t. It is given by:

$$\mathfrak{a} = \frac{\nu}{t} \tag{3.5}$$

Head-Floor distance (hd): It measures the distance between the head keyjoint coordinate & the floor where tracked person's location found.

Head-Depth Distance(d): The distance measured from first camera view to the adjacent object is termed as depth. So, head-depth is calculated via the head keyjoint's coordinates in the z-dimension of a tracked person within the frame.

Width (w)**:** Width is defined as the difference between maximum of right-left keyjoint $(\boldsymbol{R_{j\ Max}} - \boldsymbol{L_{j\ Max}})$ coordinates of tracked person. The extreme left keyjoint width is

estimated using a left-elbow, left-hip, left-knee, left-shoulder, left-ankle, left-foot, and head keyjoint values. In the same way, the right extreme keyjoints can be calculated by using all the right-side keyjoint coordinates in which R_j indicates the right keyjoint coordinates, L_j indicates the left keyjoint coordinates. It is calculated as:

$$\boldsymbol{W} = \left| \boldsymbol{R_{j\ Max}} \ - \ \boldsymbol{L_{j\ Max}} \right| \tag{3.6}$$

Height (H)***:*** Height is the measure between utmost bottom keyjoints and utmost top keyjoints of body coordinates. In extreme bottom, it includes keyjoint coordinates like left-ankle, right-knee, left-knee, right-ankle, right-foot, left-ankle, left-foot and right-ankle and in utmost top, it includes keyjoint coordinates like head, right-ankle, right-elbow, left-ankle, left-elbow, right-knee, and left-knee keyjoints coordinates, in which T_j indicates the top keyjoint coordinates,B_j indicates the bottom keyjoint coordinates.

$$\boldsymbol{H} = |\boldsymbol{T_j} \ - \ \boldsymbol{B_j}\ | \tag{3.7}$$

3.3 ConvST-LSTM: The Proposed Model

The final pre-processed 3D keyjoint coordinates are inputted into the proposed DL network. We have used the sequential fusion of CNNs, Conv-LSTM & ST-LSTM to propose the ConvST-LSTM network.

3.3.1 Convolutional Neural Network Architecture

Initially, the human action recognition has been executed by applying CNNs approach [137]. Consider, $X_t^0 = [X_1, X_2, X_3, \ldots\ldots\ldots\ldots, X_n]$ as the input-vector, where n indicates the input samples and output of convolutional layers can be defined as follows:

$$C_i^{l,j} = \sigma\left(B_j + \sum_{m=1}^{M} W_m^j * X_{i+m-1}^{0,j}\right) \tag{3.8}$$

here l corresponds to an index of convolutional layer; σ depicts the non-linear sigmoid-activation function; whereas B represents the bias vector corresponds to j^{th} feature-map; Filter-size of CNN is indicated by M; indicates the weight metrics for the j^{th} feature-map is indicated by W_m^j ; m^{th} is the filter-index.

The input frames in the proposed model consist of three input channels, namely sequences, keyjoint, and coordinates which resemble to the x, y, and z directions, respectively. Each input frame has a resolution of 125x25x3 pixels and contains infor-

mation about the movement sequences, keyjoint positions, and spatial coordinates. In convolution layer, 6 filters are passed together with configured size of kernel, padding and SoftMax functions in the hidden layer in order to avoid the vanishing gradient problem. Max-pooling is used as a pooling-operation to estimate the maximum value for feature-map, and diminish the processing time by reducing the dimensionality of the frame. Then, output from the hidden layer has passed to FC-dense layers. Finally, SoftMax function shows the prediction score of the action classes.

3.3.2 Spatio-Temporal LSTM

Before moving to ST-LSTM, let's recap LSTM [138], which consists of 3 memory cells (gates) and escape the vanishing gradient issue. These are: (a) Forget cell: a binary gate that decides how much information to pass through. (b) Input cell: decides whether the current information can be stored in the unit cell and (c) Output cell: contains sigmoid activation gate which decides which information to show as output. Lastly, the tanh layer is used to pass the cell state and further multiply it with the final output obtained from the output cell. The equations which define the activity of each cell can be formulated as follows:

$$i_t = \sigma\ (\ W_{X_i}X_t + W_{H_i}\ H_{t-1} + W_{C_i}C_{t-1} + B_i) \tag{3.9}$$

$$f_t = \sigma\ (\ W_{X_f}X_t + W_{H_f}H_{t-1} + W_{C_f}C_{t-1} + B_f) \tag{3.10}$$

$$o_t = \sigma\ (\ W_{X_o}X_t + W_{H_o}H_{t-1} + W_{C_f}C_t + B_o) \tag{3.11}$$

$$C_t = f_t C_{t-1} + i_t \tan h(W_{X_c}X_t + W_{H_c}H_{t-1} + B_c) \tag{3.12}$$

$$H_t = o_t \tan h(C_t) \tag{3.13}$$

here, W_i, W_f, W_o indicates weight matrices of forget (f), input (i) and output (o) gates, respectively; $X_t \in$ input fed to LSTM cells unit at t time; σ depicts the sigmoid-activation function whereas $tanh$ depicts the hyperbolic-tangent function (both non-linear functions); C_t indicates memory cell state within the LSTM. B_i, B_f, B_o, and B_c indicates the bias vectors on forget, input & output gates, and memory cell c, respectively. Internal frame input keyjoint coordinates of each cell in the ST-LSTM model are represented in Fig. 3.4 The skeletal keyjoints are arranged in spatial direction and input as a chain whereas the corresponding keyjoints are inputted over various frames for temporal direction sequentially. Especially, each ST-LSTM cell is feed for a new input $(x_{j,t})$, where $x \in$ new input feed for 3D position of body keyjoint j in frame time t, the hidden layer $(h_{j,t-1})$ of the same keyjoint j and the hidden layer $(h_{j-1,t})$ for the

previous keyjoint $j-1$ in same frame t, here j indicates the indices of keyjoint, i.e., $j \in 1..j..J$ and t indicates the indices of frames, i.e., $t \in 1,..t..,T$,. An ST-LSTM unit cell consists of an input cell $(i_{j,t})$, 2-forget cells correspond to the sources of context information i.e., temporal dimension $(f^T_{j,t})$ & spatial domain $(f^S_{j,t})$, in conjunction with an output gate $(o_{j,t})$. The equations for ST-LSTM are formulated as introduced in [139]:

$$\begin{pmatrix} i_{j,t} \\ f^{(S)}_{j,t} \\ f^{(T)}_{j,t} \\ f_{j,t} \\ o_{j,t} \\ u_{j,t} \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \left(W \begin{pmatrix} x_{j,t} \\ h_{j-1,t} \\ h_{j,t-1} \end{pmatrix} \right) \tag{3.14}$$

$$C_{j,t} = i_{j,t} \odot u_{j,t} + f^{(S)}_{j,t} \odot c_{j-1,t} + f^{(T)}_{j,t} \odot c_{j,t-1} \tag{3.15}$$

$$h_{j,t} = o_{j,t} \odot tanh\,(c_{j,t}) \tag{3.16}$$

where $c_{j,t}$ indicates the cell state; $h_{j,t}$ indicates the hidden input layer in ST-LSTM unit at the spatio-temporal steps for keyjoint j and frame time t; the modulated input frame is indicated by $u_{j,t}$; $\odot$ represents frame-wise product for each unit and W indicates an affine transformation within the weight.

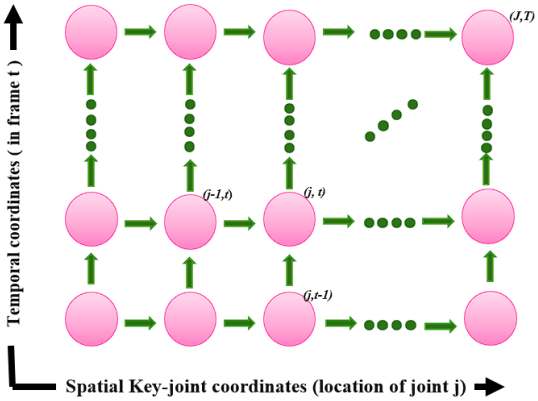


Fig. 3.4: Illustration of the ST-LSTM cell. In the spatial domain, skeletal keyjoints in each frame are aligned and fed sequentially. In the temporal domain, the keyjoints are fed sequentially across frames.

3.3.3 ConvST-LSTM-Net Architecture

Several works [140, 141] have demonstrated that each action sequence has a subset of informative keyjoints. In contrast, some keyjoints may be irrelevant in order to recognise the action classes with proper information. Therefore, for obtaining high accuracy in human-action recognition the informative skeletal keyjoints have been identified while focusing on their features vector. At the same time in order to recognize human behaviour, we must preferentially concentrate on the informative keyjoints (coordinates for feed); ignoring the features of the irrelevant keyjoints. This model has been executed by taking a sequential fusion of CNN, ST-LSTM, and FC layers. Here, CNNs are pre-owned for feature extraction, ST-LSTMs are used in sequence prediction for spatio-temporal feature extraction and the features dense layers are used for mapping. For classification, the outputs from CNN's hidden layer are fed to the ST-LSTM layers and then Global Average Pooling (GAP) layer is used to flatten the data followed by FC layers within the model. The transformation equations for ConvST-LSTM-Net can be given as:

$$\mathcal{F}_{j,t}^{(T)} = \sigma\ (\ W_{X_{\mathcal{F}}} X_{j,t} + W_{H_{\mathcal{F}}}\ H_{j,t-1} + B_{\mathcal{F}}) \tag{3.17}$$

$$\mathcal{F}_{j,t}^{(S)} = \sigma\ (W_{X_{\mathcal{F}}} X_{j,t} + W_{H_{\mathcal{F}}} H_{j,t-1} + B_{\mathcal{F}}) \tag{3.18}$$

$$\tilde{I}_{j,t} = \sigma(W_{X_{\tilde{I}}} X_{j,t} + W_{H_{\tilde{I}}} H_{j,t-1} + B_{\tilde{I}}) \tag{3.19}$$

$$\tilde{O}_{j,t} = \sigma(W_{X_{\tilde{O}}} X_t + W_{H_{\tilde{O}}} H_{t-1} + B_o) \tag{3.20}$$

$$C_{j,t} = f_{j,t} C_{t-1} + i_{j,t} \tan h(W_{X_c} X_t + W_{H_c} H_{j,t-1} + B_c \tag{3.21}$$

$$u_{j,t} = \tan h(W_{X_u} * X_t + W_{H_u} H_{t-1} + B_u \tag{3.22}$$

$$H_t = o_t\ \odot tanh(Ct) \tag{3.23}$$

where $X_{j,t}$, $C_{j,t}$, $H_{j,t}$, $F_{j,t}$, $I_{j,t}$ indicates inputs states, cells states, hidden states, forget cells, input cells for keyjoint j in frame time t; u_t input-modulation gates and $\tilde{O}_t$ is the output cells; C_t is the memory cell used for aggregating the states information controlled by the cells. Fig. 3.5 depicts about the ST-LSTM layer for each unit cell.

Model Training: ConvST-LSTM-Net was trained on the frame samples obtained from the videos, key-point recognition, followed by the fusion of Spatio-Temporal model consisting of ConvLSTMs. The model was trained on 150 epochs on a machine having AMDRyzen7 5800H processor, 8 GB RAM, and Graphics: NVIDIA GeForce RTX 3050M, GPU, having a learning rate of 0.001 after repeated hyper-parameter tuning.

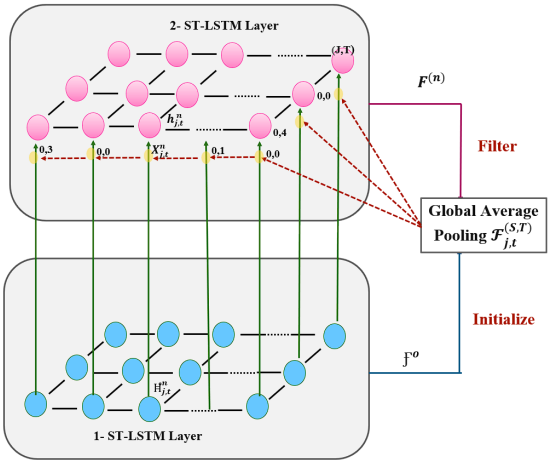


Fig. 3.5: Illustrates the ConvST-LSTM network for ST-LSTM layer in each unit cell.

For setup, Keras API version 2.3 of Python along with TensorFlow version 2.3.0 has been used in the backend to build the spatio-temporal model. To increase the code's reusability and readability, some helper functions are initially defined from the python libraries. Along with an optimum value has been set for the user-defined hyperparameters like size, no. of layers, iteration, epochs, no. of batch sizes, and learning rate. The training sample data with various batch size is feed to the model and get trained over 150 epochs. In first time-distributed CNN layer, we use 32 filters with kernel size 3 and its output is then regularized to attain faster convergence. Then, the max-pooling is added to diminish the computational cost. Dropout layer benefited to avoid the overfitting where 50% of weights are dropped randomly. For next time-distributed CNN layers, different size of filters is practices to execute feature extraction followed by another dropout layer of 40%. At step 3, we use GAP layer through which the output of CNN layer is flattened to 1*56 dimension. Further, ST-LSTM is used to handle the sequential action data of the tracked person's keyjoints coordinates. The ST-LSTM layer's output is passed to the time-distributed FC dense layer. At last, SoftMax layer gives the framewise probabilities for each action classes. The architecture of the proposed ConvST-LSTM model is illustrated in Fig. 3.6 Further, Adam optimizer is help in optimizing the cost function and uses gradient clipping within the code. The hyperparameters such as checkpoint-path, saver-function, epochs, iterations, filter size, kernel size, and test & train data have been set for training purpose. It has been found that the performance is excessively upgraded by using a sequential way.

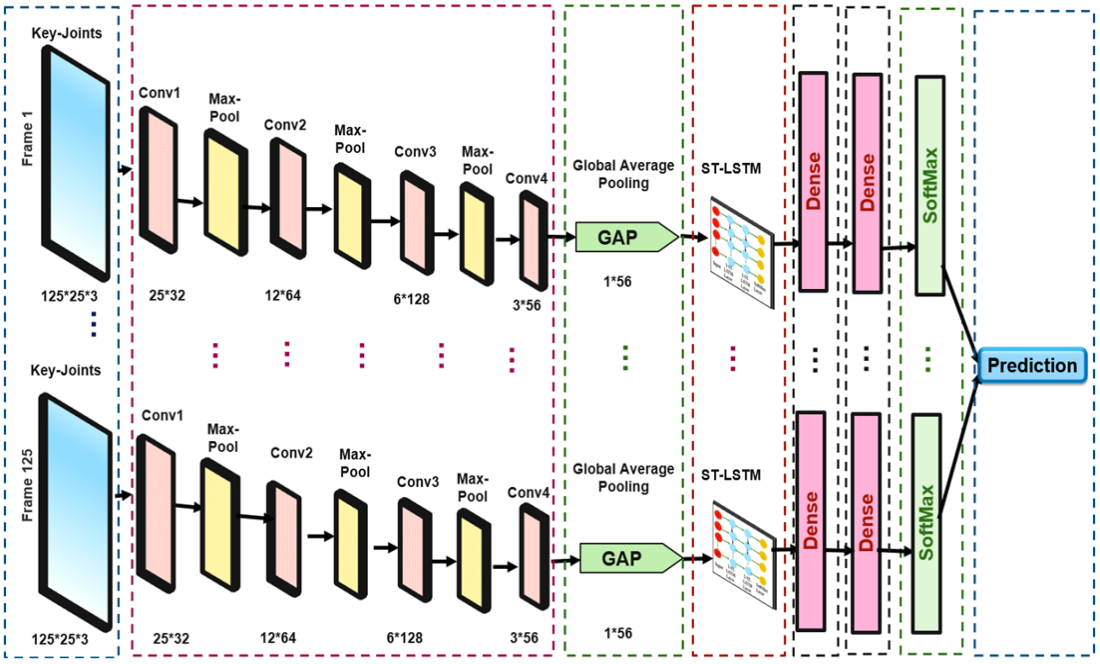


Fig. 3.6: Block architectural diagram of ConvST-LSTM-Net model for human action recognition. Starting from left side, the input frames clipped from videos; time-distributed convolutional layers including max-pooling, GAP, ST-LSTMs, FCL dense layer followed by SoftMax function layer that results as a prediction of action.

3.4 Experimental Results and Analysis

This section discusses the implementation trait of the proposed model on various benchmarks with its training hyperparameters.

3.4.1 Experiments on NTU RGB+D 60 Dataset

The NTU-RGB+D60 [113] is publicly available dataset used for HAR consisting of total 56,880 samples having 60 activity classes collected over 40 subjects. In this dataset, activities are classified into three categories having 40 daily living activities (drinking, standing, reading, happing, etc.), 9 medical conditions-related activities (sneezing, staggering, falling, vomiting etc.) and 11 common activities (punching, kicking, hugging etc.) based on multi-modal information of the daily action characterization, along with 3D skeletal keyjoint, RGB-videos, masked-depth maps, full-depth maps and infrared sequences data. The annotations provide the 3D location in x, y, z-dimension of each keyjoint in the camera coordinate system. It has total 25 key points per subject and each clip has 2 subjects. The evaluation has done on two protocols: Cross-Subject (CS) and Cross-View (CV). For performing experiments, we choose 5 action classes

(Stand, Sit, Run, Walk, Fall) contains 150 clips in each class. The two benchmarks for evaluation are set as: 1) CS contains 400 clips from 5 subjects, used for training; and the 100 clips for validation. 2) CV contains 450 from 5 subjects used for training and 150 clips for validation. The proposed ConvST-LSTM-Net model surpasses the ConvLSTM network in [137] by 4.3% with the CS evaluation protocol and 3.1% with the CV evaluation protocol. This demonstrates that spatio-temporal skeleton-based recognition approaches in LSTM networks bring significant improvement. The comparative analysis for the results of the proposed ConvST-LSTM-Net model with SOTA approaches has been enumerated in Table 3.2.

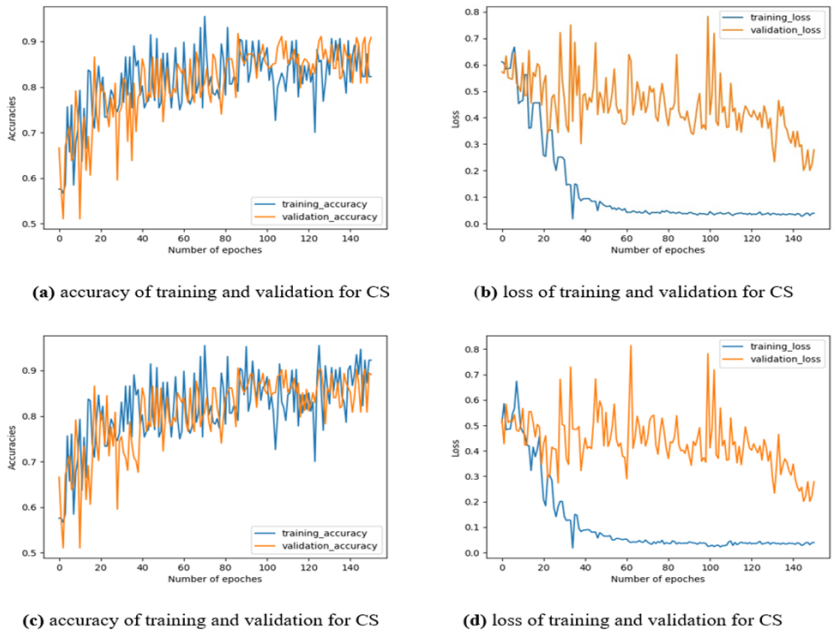
Table 3.2: Experimental Results on NTU RGB+D 60 for skeletal sequence data.

Methods	**CS %**	**CV %**
Deep-LSTM [142]	56.3	64.1
ST-LSTM [139]	69.2	77.7
ST-LSTM + Global(1) [143]	70.5	79.5
ST-LSTM + Global(2) [143]	70.7	79.4
Conv-LSTM [137]	76.2	83.2
Conv-GRU [144]	88.9	90.1
LA-GCN [145]	90.9	89.28
TD-GCN [146]	91.82	94.2
SkeletonGCL [147]	89.2	90.3
ConvST-LSTM-Net	**91.72**	**90.5**

The trade-off curves for training accuracy & loss and validation accuracy & loss on the benchmark of NTU RGB+D 60 dataset for its two-evaluation protocol i.e., CS and CV has been illustrated in Fig. 3.7 Training and validation accuracy increases with time as shown in Fig. 3.7 (a) & 3.7 (c) and finally, the growth rate reaches a steady-state value. The loss curve is shown in Fig. 3.7 (b) & 3.7 (d) which demonstrates how the validation loss gradually decreases by increasing epoch. For testing, the weights from the epochs with the maximum validation accuracy are saved.

3.4.2 Experiments on UT-Kinect Dataset

The UT-Kinect dataset [148] is publicly available and was taken through a single stationary Kinect comprised of total 10 subjects that took total 10 action types (walking, stand up, pick up, carry, sit down, throw, push, pull, wave hands, clap hands). Each subject performs each action twice. 3-channels were captured for (i) RGB, (ii) depth,



(a) accuracy of training and validation for CS

(b) loss of training and validation for CS

(c) accuracy of training and validation for CS

(d) loss of training and validation for CS

Fig. 3.7: Trade-off curves for model's Training and Validation Accuracy vs. Training and Validation Loss on the NTU RGB+D 60 benchmark dataset.

and (iii) skeleton keyjoint locations. We have only recorded the frames when the skeleton of human body was tracked. To assess the proposed method on this dataset, the standard leave-one-out-cross-validation protocol has been followed. Table 3.3 provides the comparative result of the proposed ConvST-LSTM-Net model with SOTA approaches. The trade-off curves for the model accuracy and loss on the UT-Kinect dataset has been illustrated in Fig. 3.8. It is observed from these curves that the proposed methodology offers exceptional accuracy during training and moderate accuracy in the validation process. For training process, the model causes low and for validation process it causes moderate loss.

Table 3.3: Experimental Results on UT-Kinect

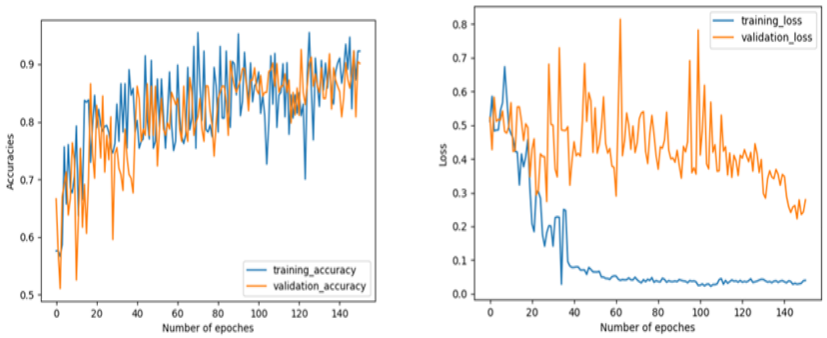
Method	**Accuracy**
Histogram of 3D Joints [148]	88.9%
ST-LSTM [139]	87.0%
ST-LSTM+Global(1) [143]	91.9%
ST-LSTM+Global(2) [143]	90.8%
Conv-LSTM [137]	90.2%
Conv-GRU [144]	89.99%
ConvST-LSTM-NET	**92.0%**

3.4.3 Experiments on UP-Fall Detection Dataset

The UP-Fall Detection Dataset [65] is the large-scale multimodal dataset collected by using vision-wearable, and ambient sensors. It includes Activity for Daily Livings (ADLs-850 GB), collected by 17 healthy persons including 9 male, 8 females individuals. It has total 11 actions i.e., 6 basic actions for daily living: walk, sit, stand, picking-up an item, laying, jump and 5 fall-actions: fall-forward via knees, fall-forward via hands, fall-sitting in an empty chair, fall backward and fall-sideward). Two cameras were set up to capture the subject's front views as well as its side views. Total 589,418 sample image frames are there taken from both cameras. Total size of this vision dataset was 277 GB. For performing experiments, we choose 5 action classes (i.e., Stand, Sit, Run, Walk, Fall) contains 1000 clips in each class in which 800 clips used for training, and the 200 clips for validation. Table 3.4 gives the comparative results of the proposed ConvST-LSTM-Net with various SOTA methods.

Table 3.4: Experimental Results on UP-Fall Detection Dataset

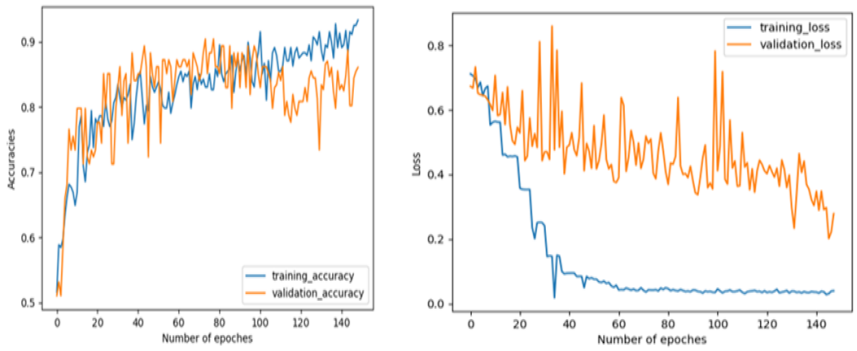
Method	**Accuracy**
GCA-LSTM [143]	88.5%
Conv-LSTM [137]	87.6%
Conv-GRU [144]	88.8%
ConvST-LSTM	**89.0%**



(a) accuracy of training and validation **(b)** loss of training and validation

Fig. 3.8: Trade-off curves for Model Training and Validation Accuracy & Model Training and Validation Loss on the UT-Kinect dataset.

Fig. 3.9 illustrates the trade-off curves for (a) accuracy of training and validation vs. (b) loss of training and validation. It is observed from these curves that the proposed



(a) accuracy of training and validation **(b)** loss of training and validation

Fig. 3.9: Trade-off curves for Model Training and Validation Accuracy vs. Model Training and Validation Loss on UP-Fall Detection dataset.

methodology offers exceptional accuracy during training and moderate accuracy in the validation process. For training process, the model causes low and for validation process it causes moderate loss.

3.4.4 Experiments on UCF101 Dataset

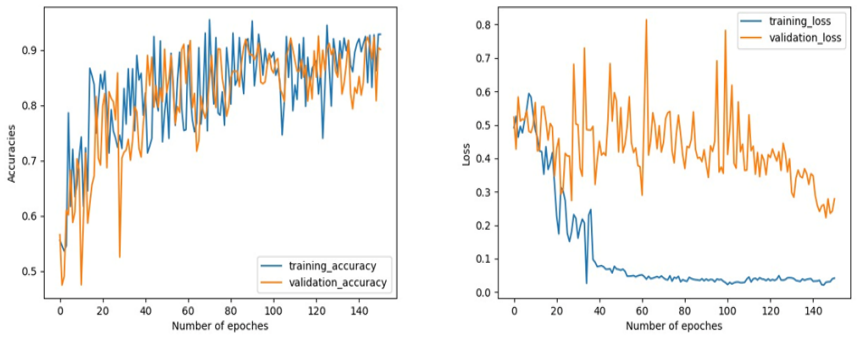
The UCF101 [149] is a popular action recognition dataset that contains 13320 video clips from 101 action categories. The action videos are clustered in 25 groups, where each group contains 4-7 videos of an action. The action categories can be classified into five distinct types: (a) Human-Object Interaction, (b) Body-Motion, (c) Human-Human Interaction, (d) Playing Musical Instruments, and (e) Sports. For performing experiments, we choose 5 action classes from body-motion categories contains total 17 body-motion clips. Table 3.5 gives the comparative results of the proposed ConvST-LSTM-Net with various SOTA methods.

Table 3.5: Experimental Results of ConvST-LSTM on the UCF101 Dataset

Method	**Accuracy**
GCA-LSTM [143]	84.2%
Conv-LSTM [137]	83.3%
Conv-GRU [144]	86.28%
PYSKL [150]	88.89%
ConvST-LSTM	**92.8%**

The trade-off curves for the model accuracy and loss on the UCF101 dataset has been illustrated in Fig. 3.10 i.e., (a) accuracy of training and validation vs. (b) loss of

training and validation. For training process, the model causes low and for validation process it causes moderate loss.



(a) Accuracy of Training and Validation **(b)** Loss of Training and Validation

Fig. 3.10: Trade-off curves for Model Training and Validation Accuracy vs. Model Training and Validation Loss on UCF101 Dataset

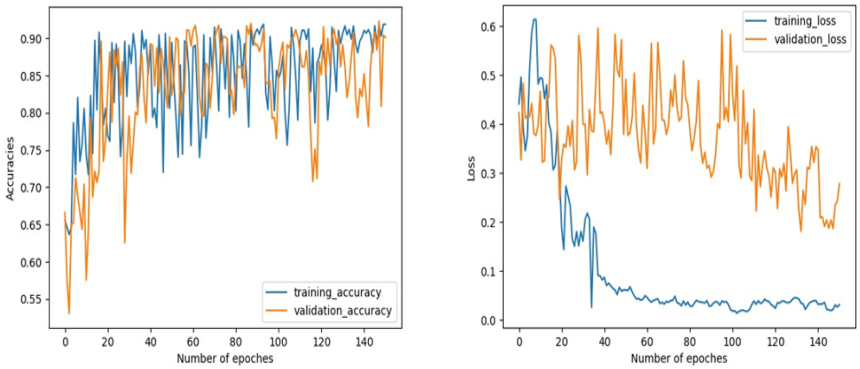
3.4.5 Experiments on HMDB51 Dataset

The HMDB51 [151] dataset is a commonly used benchmark dataset for action recognition in videos which consists of video clips from various sources like movies, YouTube. From 2GB, total 7,000 clips distributed in 51 action classes. The actions categories can be divided into five types: (a) General facial actions (b) Facial actions with object manipulation (c) General body movements. (d) Body movements with object interaction (e) Body movements for human interaction. The video clips have varying durations and resolutions. For performing experiments, we select the general body movements action classes in which 5 action clips are taken (i.e., Stand up, Sit down, Run, Walk, Fall). Each action classes contains minimum of 101 clips. Among them 80% are used of training and 20% are used for validation. Table 3.6 gives the comparative results of the proposed ConvST-LSTM-Net with various methods.

Table 3.6: Experimental Results on the HMDB51 Dataset

Method	**Accuracy**
GCA-LSTM [143]	82.3%
Conv-LSTM [137]	81.2%
Conv-GRU [144]	80.8%
PYSKL [150]	69.4%
ConvST-LSTM	**91.86%**

Fig. 3.11 illustrates the trade-off curves for (a) accuracy of training and validation vs. (b) loss of training and validation on HDMB51 dataset. Form this, it is observed that the proposed methodology achieves outstanding accuracy during the training process and moderate accuracy during the validation process. The model exhibits low loss during training and moderate loss during validation.



(a) Accuracy of Training and Validation **(b)** Loss of Training and Validation

Fig. 3.11: Trade-off curves for Model Training and Validation Accuracy vs. Model Training and Validation Loss on HDMB51 Dataset.

3.4.6 Multimodal Analysis over Standard Performance Measures

The performance of the proposed model has been measured on different performance metrics. Fig. 3.12 represents the accuracy, precision, recall, and F1-score scores on different benchmarks. The accuracies and losses are plotted for 150 epochs. The proposed ConvST-LSTM-Net results in a better accuracy. Fig. 3.13 illustrates the human action recognition results obtained in different benchmarks datasets with framing the bounding box over the tracked human. We observed that the performance of the model is sufficiently high.

3.5 Summary

In this chapter, we presented a novel deep learning-based framework for skeleton-based HAR, termed as ConvST-LSTM-Net. The proposed model aims to effectively capture both spatial and temporal features of human skeletal data by integrating CNN with ST-LSTM units. The motivation behind this model stems from the need to identify and utilize only the informative keyjoints within human skeleton sequences, as irrelevant keyjoints often introduce noise and degrade performance in conventional meth-

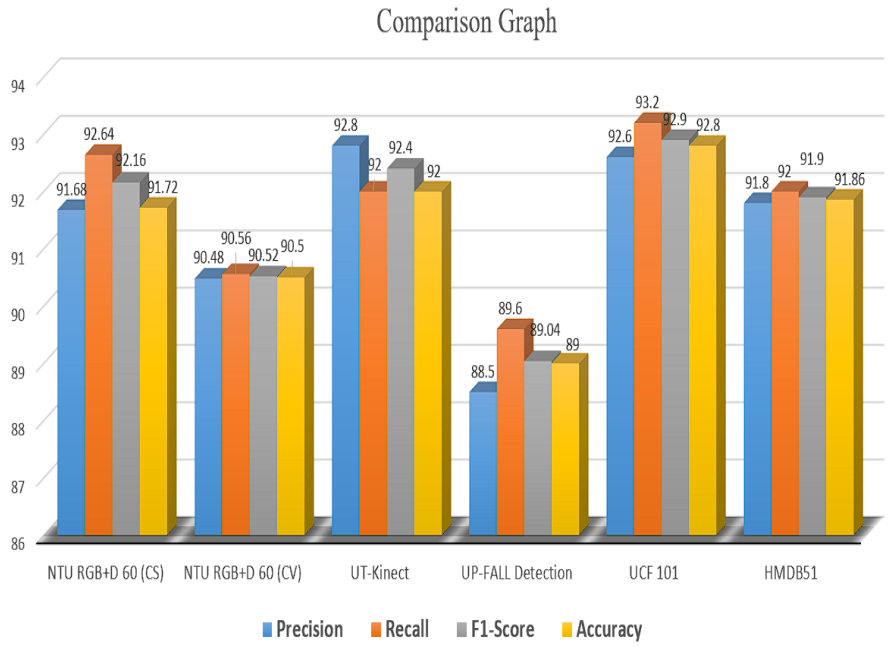


Fig. 3.12: Comparative Stats of Standard Performance Measure over different datasets.

ods. The ConvST-LSTM-Net begins with the detection and preprocessing of skeleton keyjoints from RGB video frames. A feature construction module is developed that extracts geometric and kinematic features such as joint angles, motion velocity, acceleration, and posture metrics, which are essential in characterizing human actions. These features are then passed through convolutional layers for spatial feature extraction, followed by ST-LSTM layers to model the temporal dynamics across video frames. Finally, the processed information is classified using fully connected dense layers and a SoftMax layer to predict the action class. The experimental results show that the proposed approach works adequately in restoring the details hidden in the dark areas. Furthermore, extensive experiments demonstrate that ConvST-LSTM-Net significantly outperforms existing models, especially in handling complex scenarios involving variations in viewpoint, background, and motion complexity. The training-validation accuracy curves and loss metrics confirm the model's robustness, with consistently high accuracy and low generalization error across diverse datasets.

Fig. 3.13: Illustration of the Human Action Recognition on various benchmarks. Starting from left–right **(a)** NTU RGB+D 60 Dataset: Sitting, Standing **(b)** UT-Kinect Dataset: Standing, Walking **(c)** UP-Fall Detection Dataset: Fall **(d)** UCF101 Dataset: Walking, Running e HMDB51 Dataset: Running

Chapter 4

STAD-ConvBi-LSTM: Spatio-Temporal Attention-based Deep Convolutional Bi-LSTM Framework for Abnormal Activity Recognition

In this chapter, we present a novel deep learning-based framework designed to identify and recognize abnormal human activities in video sequences. The proposed approach integrates convolutional neural networks, bi-directional LSTMs, and spatial-temporal attention mechanisms to enhance feature discrimination and long-term pattern recognition.

4.1 Introduction

HAR is an area that analyzes the hidden consecutive pattern of human activity and predicts its state of action, whether it is normal or abnormal, based on perceptual situations, as in the video frames. In contrast to images, video contains more information. In video, a grouping of different human parts in motion is termed as human activity, likewise body+hand+ arms+legs+face or a grouping of all body parts movements. Formerly, the research based on abnormal detection of human patterns was entirely attentive to the activities performed by a single or multiple humans, including a single object, but in a controlled setting [152].

Recently, DL-based approaches have embraced the integration of RNN to more effectively understand and recognize complex human activities in video sequences instead of single frames. Generally, CNN is subjected to extracting differential feature information, while RNN is specifically subjected to gaining knowledge of hidden sequential patterns within these extracted CNN features from videos. Most of the existing HAR approaches have been implemented on large pre-trained CNN models trained on image datasets.

Therefore, we present a framework, "STAD-ConvBi-LSTM", that focuses on gathering spatial-temporal feature representation. Specific emphasis on distinctive feature learning in the long-term video sequence to identify abnormal activity within the frame. The main focus is on extracting the key distinguishing information from the video sequence & utilizing its updated attention weights for activity prediction. The primary contribution of the proposed STAD-ConvBi-LSTM framework lies in employing a CNN with remaining attention blocks to enhance the features. We propose a feature fusion-based framework of bi-LSTM with a channel-wise Spatial-Temporal Attention (STA) mechanism for a more comprehensive understanding of the spatial-temporal representations within sequential information. The attention weight is dynamically adjusted based on learned global features to identify human activities in a sequence effectively. Fig. 4.1 depicts the graphical workflow of the STAD-ConvBi-LSTM framework. To tackle the limitations of existing approaches, we presented a DL-based efficient framework for HAR challenges that recognizes the abnormal activities in terms of channel-wise spatial-temporal feature vectors extracted from video sequences. The key contributions are summarized as follows:

- We proposed a new framework by integrating deep-learning methods to identify and recognize the abnormal human activity. The model uses the prominent discriminative RGB features by combining a CNN architecture for prominent discriminative feature extraction, a bi-directional LSTM for capturing long-term modeling, and a channel-wise spatial-temporal attention mechanism to emphasize particular spatial-temporal features in unprocessed video streams.

- We propose a dual channel-wise RGB and spatial-temporal attention mechanism, which is sited after each two successive spatio-temporal convolutional layers for focusing on specific abnormal activity regions in video sequences.

- Also, introduced a bi-directional LSTM net having 6 layers (3 forward + 3 backward) that efficiently capture long-term temporal patterns of abnormal human activities, which significantly improves features' reusability.

- We aimed an ablation study to assess the advantages of proposed model to enhance the recognition accuracy. The experiments are demonstrated on publicly available datasets, i.e., UCF50 [153], YouTube Action [154], HMDB51 [151], UCF101 [149], Kinetics-600 [155] and on our synthesized dataset to show the efficiency and aptness of the proposed framework. The model shows a consistent improvement in accuracy as compared to SOTA methods.

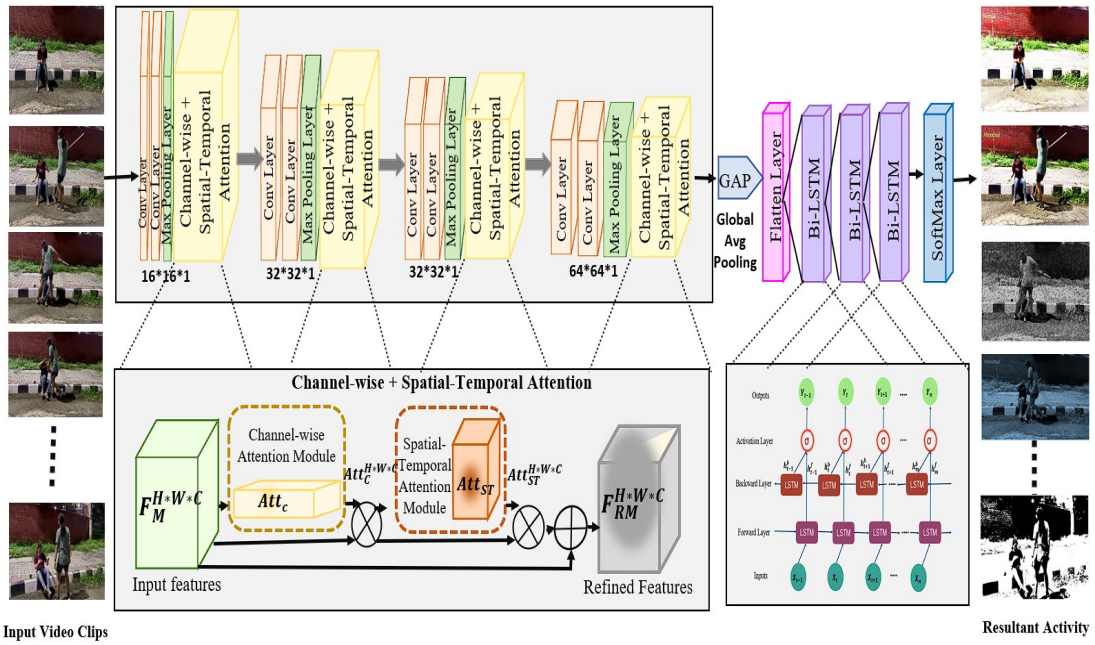


Fig. 4.1: Outline of proposed framework for abnormal activity recognition. The designed framework comprises of three components: CNN architecture, dual channel-wise spatial-temporal attention module, and bidirectional-LSTM net.

4.2 STAD-ConvBi-LSTM: The Proposed Methodology

The section describes our proposed deep framework STAD-ConvBi-LSTM as well as its core components. The framework is alienated into three modules: A, B, and C. Component A presents a lightweight CNN architecture. Component B comprises dual attention, incorporating a channel-wise STA module. This module is integrated into the CNN module, enhancing our CNN architecture in component A with a dual-attention mechanism for extracting the relevant features from video sequences. Lastly, component C consists of bi-LSTM net use for learning long-term encoded human activity.

4.2.1 Component A: CNN Architecture Module

Identifying abnormal human activity within video streaming data remains a hectic and challenging task. In this, video frames represent the complex activities of humans over a series of mounted frames as per the distinct hidden visual contents. It includes the spatial-temporal flow of human abnormal activities within the video frames, changing textures, colors, and backgrounds. These video visuals need to be examined precisely and effectively for a more in-depth and clear representation of abnormal human activities. Thus, CNN-based techniques are widely used to extract these unusual features

from these hidden activities within the video's frames efficiently. Due to the large network architectures of the CNN-based method, its computational complexity and runtime execution are still very high; instead, it shows remarkable performance. Thus, we projected a lightweight CNN architecture-module composite of spatial-temporal with an attention mechanism to overcome such limitations. The component A consists of 12 Conv layers: the 2-successive convo layers with respect to the max pooling layer followed by the STA block. Initially, 2nd convolutional layers with 8 kernels were applied to each input frame, each having kernel size 3*3. Subsequently, a total of 32 kernels each applied to the 3rd and 4th convolutional layer to the outcome of 1st dual attention block having a 3*3 kernel-size. Likewise, 5th & 6th convolutional layers utilize 32 kernels each on the outcome of 3rd dual attention block having a 3*3 kernel-size. The final pair of spatial-temporal convolutional layers applied 64 kernels each to the outcomes of the 3rd dual attention block having a 3*3 kernel size. They subsequently transmitted the evaluated feature vector to the final dual attention block. Finally, outcome from the preceding dual attention block undergoes processing through a GAP. Subsequently, a flattened layer was used to flatten the outcome. Then, the processed output is integrated with a bi-LSTM net, facilitating subsequent long, short-term sequence learning of human activity. Notably, we employed a maximum of 64 convolutional kernels/layers, maintaining a consistent kernel size of 3*3. This strategic choice significantly contributes to minimizing the computational complexity of the model while having a slight impact on its performance.

4.2.2 Component B: Dual-Attention Module

We introduce a CNN architecture driven by attention mechanisms to feature the relevant regions and improve feature representation effectively. In this component, the dual-attention mechanism is proposed using a convolutional block attention module (CBAM) introduced by Sanghyun et al. [156] using a 3*3 kernel size convo-layer, followed by the fusion of the STA model. The fusion helps diminish the overall parameter overhead. Fig. 4.2 depicts the overall architecture of the Dual-Attention Models, i.e., Component B.

The channel-wise attention module evaluates the weighted input of RGB channels-wise via employing the middle channel attention A_C on the outcome F_M of the preceding convolutional layer for achieving the channel attention Att_C as given in Eq.4.1. The outcome obtained after Att_C is after the STA block, which prominence the abnormal activities area by employing it to calculate channel feature maps. Lastly, the resultant refined feature representation F_{RM} are attained by fusing the $Att_{ST}^{H\times W\times C}$ STA

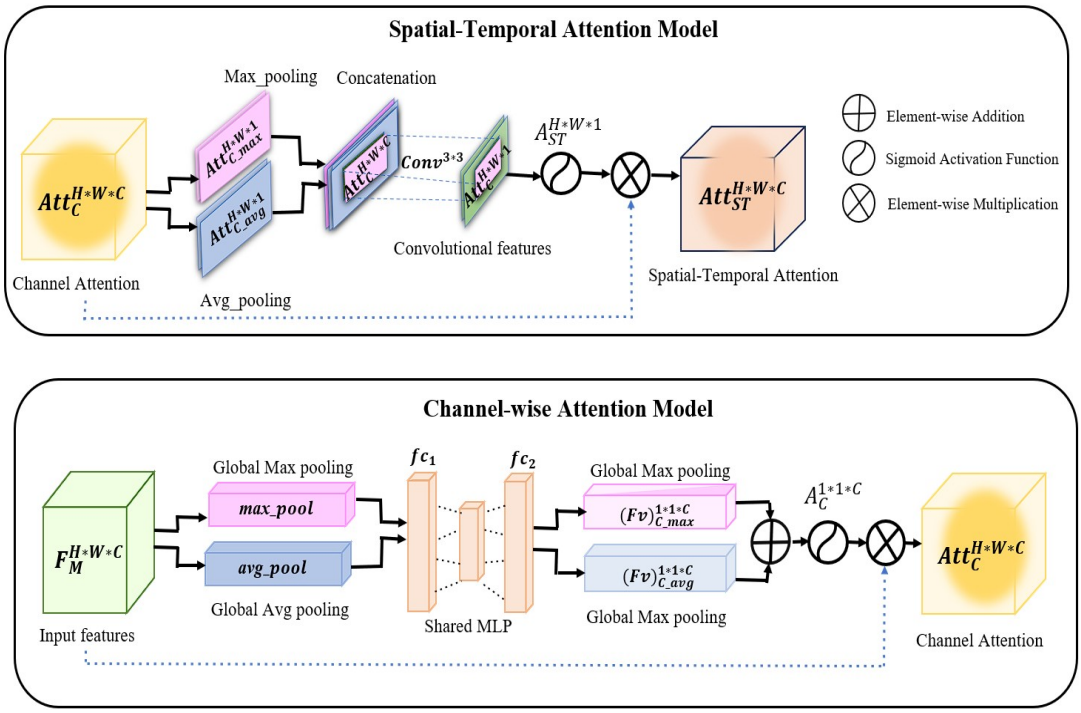


Fig. 4.2: Overview of the Dual-Attention Module consists of Channel-wise and Spatial-Temporal Attention Models

feature maps and $F_M^{H\times W\times C}$ input feature map as given in Eq. 4.4 and Eq. 4.5, respectively. Mathematically, Att_C, Att_S,Att_T and F_{RM} can be formulated as:

$$\text{Att}_C^{H\times W\times C} = \text{A}_C\left(F_M^{H\times W\times C}\right) \otimes F_M^{H\times W\times C} \tag{4.1}$$

$$\text{Att}_S^{H\times W\times C} = \text{A}_S\left(\text{Att}_C^{H\times W\times C}\right) \otimes \text{Att}_C^{H\times W\times C} \tag{4.2}$$

$$\text{Att}_T^{H\times W\times C} = A_T\left(\text{Att}_C^{H\times W\times C}\right) \otimes \text{Att}_T^{H\times W\times C} \tag{4.3}$$

$$\text{Att}_{ST}^{H\times W\times C} = \text{Att}_T^{H\times W\times C} + \text{Att}_S^{H\times W\times C} \tag{4.4}$$

$$F_{RM}^{H\times W\times C} = \text{Att}_C^{H\times W\times C} + \text{Att}_{ST}^{H\times W\times C} + F_M^{H\times W\times C} \tag{4.5}$$

where A_C, A_S and A_T are the channel-attention, spatial-attention and temporal-attention, respectively. F_{RM} indicates final refined feature vectors attained via spatial attention along with the input feature vector.

4.2.2.1 Channel Attention

In HAR, for recognizing the human pattern, each color channel of RGB provides a different pattern, as per the presence of color in the image. The Convo layer's filter operates within the local receptive field, and the feature channels $F_M^{H\times W\times C}$ are inde-

pendent. Throughout the training phase, extraction of the feature vector from the input image is done by the CNN model, which is completed by the totality of convolution layers used, in which each channel gives more feature representation than other channels. GAP and GMP are used for evaluating equally weighted feature vectors for each channel and opting for a highly activated maximum value of F_M from the receptive field, respectively. Further, these feature vectors are shared with MLP encompassing two FCL, i.e., fc_1 and fc_2, having 128 & 512 learning nodes, respectively. Both FCs operate on each point, whereas MLP is a multi-layer perceptron on each end. A shared MLP means we are applying the same MLP to each point in the point cloud. It learns the non-linearity amongst fc_1 and fc_2 FC layers via the ReLU function and produces two distinct feature vectors viz. $(Fv)^{1\times1\times C}_{C\text{-max}}$ and $(Fv)^{1\times1\times C}_{C\text{-avg}}$ for Global Map Pooling (GMP) and GAP, as given in Eq.4.6 and Eq.4.7, correspondingly. Subsequently, the evaluated feature vectors (Fv) are cumulative through an element-wise addition operation. Thereafter, it passed through the activation function to normalize and produce an intermediate channel attention learned feature, i.e., $A^{1\times1\times C}_C$ as obtained by Eq.4.8. Then, the attained middle spatial-temporal features are integrated with feature vector input $F^{H\times W\times C}_M$ via element-wise multiplying operation. Lastly, the resultant channel attention feature maps are obtained, i.e., $Att^{H\times W\times C}_C$ as given in Eq.4.9.

$$(Fv)^{1\times1\times C}_{C\text{-max}} = fc_2\left(\text{ReLU}\left(fc_1\left(\text{max\_pool}\left(F^{H\times W\times C}_M\right)\right)\right)\right) \tag{4.6}$$

$$(Fv)^{1\times1\times C}_{C\text{-avg}} = fc_2\left(\text{ReLU}\left(fc_1\left(\text{avg\_pool}\left(F^{H\times W\times C}_M\right)\right)\right)\right) \tag{4.7}$$

$$A^{1\times1\times C}_C = \sigma\left((Fv)^{1\times1\times C}_{C\text{-max}} \oplus (Fv)^{1\times1\times C}_{C\text{-avg}}\right) \tag{4.8}$$

$$\text{Att}^{H\times W\times C}_C = A^{1\times1\times C}_C \otimes F^{H\times W\times C}_M \tag{4.9}$$

where $(Fv)^{1\times1\times C}_{C\text{-max}}$ and $(Fv)^{1\times1\times C}_{C\text{-avg}}$ represent the calculated feature vectors obtained from GMP and GAP operations, respectively, $F^{H\times W\times C}_M$ signifies input feature-maps; sigmoid activation function is symbolized by σ, and $\text{Att}^{H\times W\times C}_C$ indicates eventual output channel attention.

4.2.2.2 Spatial-Temporal Attention (STA)

Our framework leverages the inter-spatial features with the temporal features to the specific areas within the frames, which aids in tracing the target human activity within the feature maps (F_M) and evaluating the features between the channels through implementing max-pooling & average-pooling to input channel-wise attention feature maps for attaining max_pool channel-attention $\text{Att}^{H\times W\times C}_{C\text{-max}}$ and average_pool channel-

attention $\text{Att}_{C\text{-avg}}^{H\times W\times C}$, respectively. Further, both the channel is concatenated subsequently to a single convo layer $Conv^{3\times3}$ with a kernel size of 3*3 pooled F_M to generate single channel convoluted F_M, i.e., $(Conv^{3\times3}\left(\text{Att}_{C\text{-max}}^{H\times W\times1}\oplus\text{Att}_{C\text{-avg}}^{H\times W\times1}\right))$ and then processed by activation function, to normalize and produces intermediate spatial-temporal attention learned feature i.e., $A_S^{H\times W\times1}$ and $A_T^{H\times W\times1}$ for single channel as given in Eq.4.12 and Eq.4.13, respectively. Then, the attained middle spatial-temporal features are integrated with channel input $Att_C^{H\times W\times C}$ using element-wise multiplication operation. Finally, the resultant spatial-temporal attention feature maps are obtained by Eq.4.14. Mathematically, spatial-temporal attention channel-wise framework with its component can be formulated as follows:

$$\text{Att}_{C\text{-max}}^{H\times W\times C}=\text{max\_pool}\left(\text{Att}_C^{H\times W\times C}\right) \tag{4.10}$$

$$\text{Att}_{C\text{-avg}}^{H\times W\times 1}=\text{avg\_pool}\left(\text{Att}_C^{H\times W\times C}\right) \tag{4.11}$$

$$A_S^{H\times W\times1}=\sigma\left(\text{Conv}^{3\times3}\left(\text{Att}_{C\text{-max}}^{H\times W\times1}\oplus\text{Att}_{C\text{-avg}}^{H\times W\times1}\right)\right) \tag{4.12}$$

$$A_T^{H\times W\times1}=\sigma\left(\text{Conv}^{3\times3}\left(\text{Att}_{C\text{-max}}^{H\times W\times1}\oplus\text{Att}_{C\text{-avg}}^{H\times W\times1}\right)\right) \tag{4.13}$$

$$\text{Att}_{ST}^{H\times W\times C}=\left(A_S^{H\times W\times1}+A_T^{H\times W\times1}\right)\otimes\text{Att}_C^{H\times W\times C} \tag{4.14}$$

where $Att_{C\text{-max}}^{H\times W\times C}$ and $Att_{C\text{-avg}}^{H\times W\times1}$ are the GMP and GAP features, correspondingly. σ indicates the sigmoid activation function; The outcomes $Att_S^{H\times W\times C}$ and $Att_T^{H\times W\times C}$ are the resultant spatial-temporal information on a single channel. $Fv\in F_M^{n\times T\times C\times H\times W}$ indicates input feature vector in which n, T, C, H, W represent batch size, temporal dimension, totality of channel, height and width respectively; Fig. 4.3 illustrates the spatial convolution $A_S^{H\times W\times1}$ with filter dimension 1*3*3 and results in spatial descriptor thru sliding across the spatial dimension H*W without altering the feature map size and also decreases the number of parameters to 1/C of traditional 3D convolution. While the temporal convolution $A_T^{H\times W\times1}$, having filter dimension 3*1*1, succeeds in learning temporal connectivity among video frames via actions and also decreases the number of channels to 1.

4.2.3 Component C: Bi-LSTM

The Bi-LSTM is a sequential model consisting of two LSTMs. The input is taken in both directions by two LSTMs. It extracts information from previous and forthcoming to make the model more time-dependent. The outcome of a bi-LSTM not only depends on prior frames but also on upcoming frames. However, LSTM has a limitation that

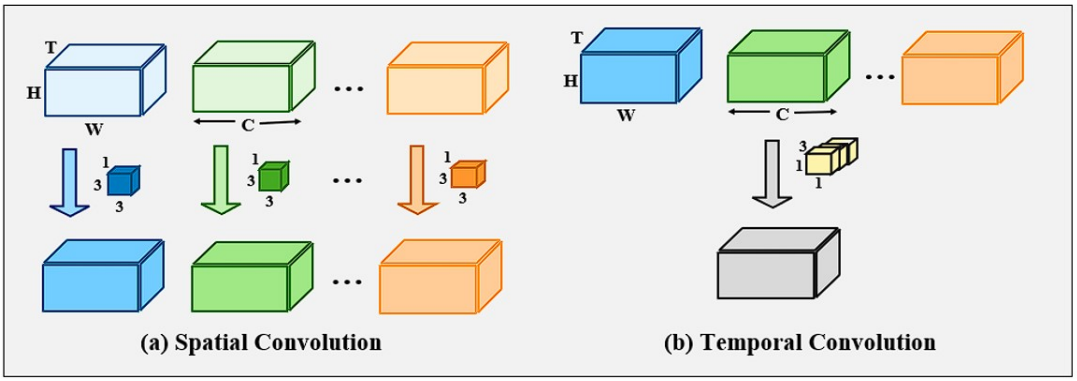


Fig. 4.3: Details of the Spatial-Temporal Convolution Layers

it only considers the prior context of input and cannot consider any future context. To address this limitation, we opted for the bi-LSTM model [60], which enables us to consider both the past and future context of the input, as depicted in Fig. 4.4 It comprises three gates, namely input, forgets and output gates. The input gate, along with the input modulation gate, keeps the new information in the memory cell at time t. Then, the forget gate regulates the information within the memory cell, whether to reject or hold the information. Finally, the output gate processes the resultant outcome of memory cell which acts as a hidden state. This committal to memory is adaptively learning the relevant information using back-propagation. This modification can convey accurate information to forthcoming predicting LSTM model for longer dependency, especially for temporal information. The evaluation of l^th hidden layer at t^th time is formulated as:

$$i_t^l = \sigma\left(W_i^l \cdot H_t^{(l-1)} + \bar{W}_i^l \cdot H_{(t-1)}^l + b_i^l\right) \tag{4.15}$$

$$f_t^l = \sigma\left(W_f^l \cdot H_t^{(l-1)} + \bar{W}_f^l \cdot H_{(t-1)}^l + b_f^l\right) \tag{4.16}$$

$$o_t^l = \sigma\left(W_o^l \cdot H_t^{(l-1)} + \bar{W}_o^l \cdot H_{(t-1)}^l + b_o^l\right) \tag{4.17}$$

$$u_t^l = \tanh\left(W_u^l \cdot H_t^{(l-1)} + \bar{W}_u^l \cdot H_{(t-1)}^l + b_u^l\right) \tag{4.18}$$

$$c_t^l = \left(f_t^l \odot c_{(t-1)}^l + i_t^l \odot u_t^l\right) \tag{4.19}$$

$$h_t^l = o_t^l \odot c_t^l \tag{4.20}$$

where i_t^l, f_t^l, o_t^l, c_t^l, and h_t^l indicate the input gate, forget gate, output gate, memory cell and hidden state, respectively. Tangent tanh and σ signify hyperbolic tangent and sigmoid activation functions, respectively. The unidirectional LSTM is inferior in performance due to its focus on retaining the information from past sequences. This

results in a limitation in capturing the relative details derived from past and future inputs and thus compromises its overall effectiveness. As the activity captured by LSTM is sequential, it makes the forthcoming information more relevant for its better prediction. Thus, bi-LSTM takes advantage of two hidden states to maintain the information at each time. The evaluation of hidden states h in bi-LSTM is defined as:

$$\overrightarrow{h}_t^l = \overrightarrow{o}_t^l \odot \overrightarrow{c}_t^l \tag{4.21}$$

$$\overleftarrow{h}_t^l = \overleftarrow{o}_t^l \odot \overleftarrow{c}_t^l \tag{4.22}$$

where $\overrightarrow{h}_t^l$ and $\overleftarrow{h}_t^l$ represent the hidden states in forward and backward directions, respectively. Both hidden states worked separately as distinct hidden layers via receiving sequence from $t = 1$ till T and vice versa. After the evaluation of both hidden states, they are fused at the hidden state as the last layer L, i.e., h_t^L the. The hidden states in the previous layer for all T segments of the trained bi-LSTM network are defined as:

$$H^L = \{h_1^L, h_2^L, \ldots, h_T^L\}; \quad t = 1, 2, 3, \ldots, T \tag{4.23}$$

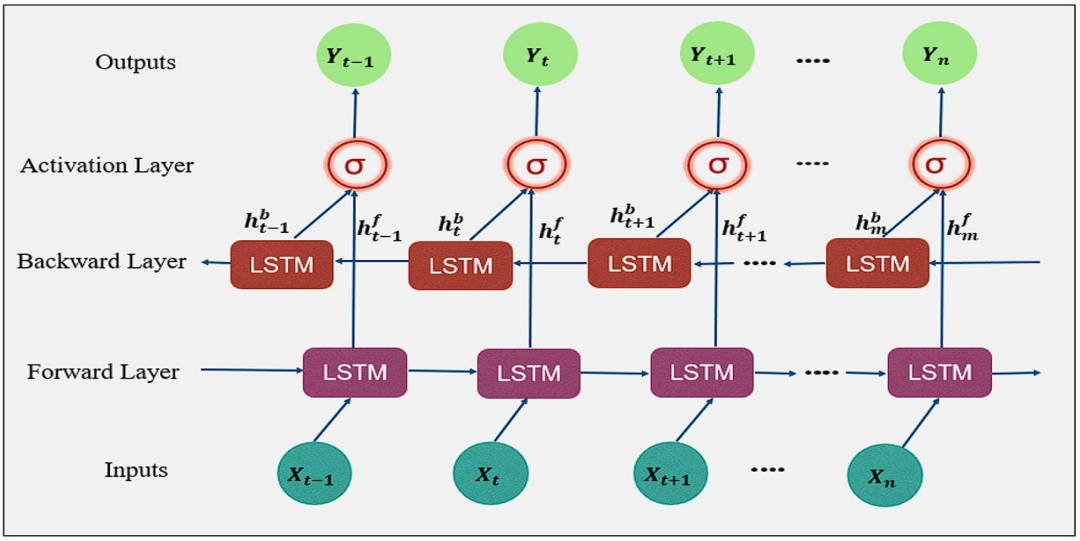


Fig. 4.4: The Architecture of Bi-directional Single LSTM Layer

4.3 Experimental Results and Performances

We conducted an extensive experiment on various benchmark datasets and on our own synthesized dataset, as elaborated below. Also, discuss the performance accuracies with SOTA.

4.3.1 Datasets

In this subsection, five benchmark datasets & our synthesized dataset are discussed.

4.3.1.1 UCF50 Dataset

The UCF50 dataset [153] poses a significant challenge as a large-scale HAR dataset. It comprises videos depicting a wide range of human activity, recorded with diverse viewpoints, camera's angle, human poses and appearances, as well as background clutter. In this, total 6,676 video clips are available, which are classified into 50 distinct activity classes and further assembled into 25 clusters. Each cluster contains five video clips of same human activity features. For performing experiments, we choose 5 activity classes (i.e., salsa, jumping, skijit, rope climbing, diving) containing 125 video clips in each class. The period of video clips ranges from 5-6 seconds having a resolution of 480*640.

4.3.1.2 UCF101 Dataset

The UCF101 [149] is a highly exciting dataset made up of 13,320 YouTube clips that are classified into a total of 101 distinct action classes. Further, each class consists of 100 video clips of five different human activities that include human-object interaction, human-human interactions, human movement, human plays with musical instruments and various sports activities performed by humans. The duration of video clips ranges from 5-6 seconds with 480*640 resolution.

4.3.1.3 HMDB51 Dataset

The HMDB51 [151] dataset is recognized as a challenging benchmark for HAR in videos and is compiled from several sources such as movies, available databases, YouTube, and Google videos, consisting of 6,849 video frames of human activities, which are classified into 51 distinct classes. Further, each class consists of 101 video clips of five different human activities: facial-actions class, facial-object actions, human-body motion, human-object interface and human-human interface classes. For performing experiments, we select normal and abnormal activity category in which 5 activity classes are taken. Each activity class contains minimum of 101 clips.

4.3.1.4 YouTube Action Dataset

The YouTube action dataset [154] recorded from YouTube contains miscellaneous sports and other action video clips. The video clips presented have some difficulties due to camera movement, cultured and complex backgrounds, variations in viewpoints, diverse poses, and the occlusion of other objects. Total of 1640 video clips are grouped into 11 human action classes with interval times of 2-5 seconds and resolution 320*240. Further, the 11 human action classes are clustered in 25 clusters comprising 4-6 video clips in a similar group sharing common film features. Out of 11 human action classes we have taken the 5 activity classes (i.e., Jumping, Jugging, basketball, diving, Swinging) with video resolution 480*640 and duration time of 2-5 seconds.

4.3.1.5 Kinetics-600 Dataset

Kinetics-600 [155] is a large-scale action recognition video dataset comprising 4,80,000 video clips with 600 action categories of different human actions. The total 480,000 videos are separated into 3 distinct sets for training, validation and testing, i.e., 390000, 30000, and 60000 videos, respectively. For performing experiments, we select the general 5 activity classes (i.e., Jogging, dancing, painting, fighting, threatening). Each video clip has interval times of 10 sec and is annotated from a raw YouTube video of resolution 480*640.

4.3.1.6 Synthesized Human Action Dataset

We are developing a synthesized human action dataset in the DTU campus and outside the campus as well; till now, out of 350, only 300 video clips have been synthesized. The video clips are recorded through a camera of 12 Megapixels, f/1.6, 26 mm wide, 1/1.9", 1.7μm with PDAF dual pixel and sensor-shift OIS at a resolution of 480*640. Currently, it contains 5 activity classes, viz. Adverse, sports, vandalism, normal, Shoplift categories. The dataset encompasses 250 video clips with interval times of 5-6 seconds, each class consist of 60 video clips. We have chosen 105 video clips as a random sample for training set, and the remaining sample used for testing. Fig. 4.5 illustrates some occurrences of video clips for the above dataset.

4.3.2 Implementation Details

The experiment is implemented on Python v3.7 using libraries such as TensorFlow, Keras API version 2.3, OpenCV, Anaconda, and CUDA 10. The proposed model is

Fig. 4.5: Instances of Synthesized Action dataset

conducted on a machine with CUDA-enabled GPU on a Windows server having configuration AMD Ryzen7 with 5800H processor, Graphics: NVIDIA GeForce RTX 3050M with 16 GB RAM. For sequence learning and optimized detection, 20 frames of sequence length are extracted from the video as input and resized by 224*224*3 resolutions with no overlap in both forward and backward directions through bi-LSTM. Meanwhile, 3 bi-LSTM layers are used with 32 bi-LSTM per layer. The dataset was partitioned into a 70-30% ratio: 70% was utilized for training, and for validation the remaining 30% data was utilized. Overall, the model is trained on 200 epochs, utilizing Adam optimization having a 0.0001 learning rate while choosing the batch sizes of 16,32,64,128, respectively. The proposed STAD-ConvBi-LSTM model operates on categorical cross-entropy loss used to reduce the biased weight tuning on model prediction during training, and a dropout of 25% rate is added to prevent overfitting. Likewise, the center loss technique is used in the model, which enhances the performance of video motion classification. This technique resolves the problem of intra-class variability. The hyperparameters are selected carefully to ensure the model's effectiveness while learning and to produce consistent results. Table 4.1 displays the selected hyperparameters of the proposed framework.

Table 4.1: Hyperparameters castoff for STAD-ConvBi-LSTM Framework

#Parameter	Merits
Resolutions Input Frame	224×224×3
No. of CNN Layers	4
Filter-Size	16, 32, 64, 128
Kernal-Size	3×3(fixed)
Global Max Pooling	Yes
Global Average Pooling	Yes
No. of Epochs	200
Batch-Size	8, 16, 32, 64
Optimizer Used	Adam Optimizer
Loss-Function Used	Categorical cross-entropy loss
Dropout-Rate	25%

4.3.3 Results Analysis

This section discusses the results analysis gained on the proposed framework as discussed below.

4.3.3.1 Ablation Studies of the Proposed Framework with Baseline Methods

The result analyses with performance comparisons with SOTA are presented individually for each dataset. The experiment revolves around exploring various possible solutions for HAR. Throughout the process, we devised some spatial-temporal HAR methods on 6 distinct baseline approaches, including CNN+LSTM having 8 Conv+3 LSTM, CNN+bi-LSTM having 8 Conv+6 LSTM for both directions (3 forward & 3 backward LSTM), CNN+GRU having 8 Conv+3 GRU, CNN+bi-GRU having 8 Conv+6 GRU for both directions (3 forward & 3 backward LSTM), DA-CNN+bi-GRU [157] having 12 Conv+6 GRU for both direction (8 Conv+ 4 Attentional and 3 forward & 3 backward LSTM) and we evaluate their effectiveness, and finally presented our proposed framework STAD-ConvBi-LSTM having 12 Conv+6 LSTM for both direction (8 Conv+ 4 Attentional and 3 forward & 3 backward LSTM). Further, the STAD-ConvBi-LSTM framework is trained on these datasets. Table 4.2 outlines the detailed system setup of these baseline approaches.

Table 4.2: Experimental System for Baseline Approaches & Proposed Framework

Approach	**Channel-wise Layer**	**Spatial Convolutional Layer**	**Temporal Convolutional Layer**
CNN + LSTM	RGB	8 Conv	3 LSTM Layers
CNN + bi-LSTM	RGB	8 Conv	6 bi-LSTM (3 Forward + 3 Backward)
CNN + GRU	RGB	8 Conv	3 GRU Layers
CNN + bi-GRU	RGB	8 Conv	6 GRU Layers (3 Forward + 3 Backward)
DA-CNN + bi-GRU	RGB	12 Conv (8 Conv + 4 Atten.)	6 GRU Layers (3 Forward + 3 Backward)
STAD-ConvBi-LSTM (Proposed)	**RGB**	**12 Conv (8 Conv + 4 Atten.)**	**6 bi-LSTM (3 Forward + 3 Backward)**

Table 4.3: Comparative Study of STAD-ConvBi-LSTM Framework besides Baseline Approaches

Approach	**Dataset**	**Acc %**
CNN+LSTM	UCF50 [153]	78.9
CNN+Bi-LSTM		82.8
CNN+GRU		84.2
CNN+Bi-GRU		88.5
CNN+Bi-GRU (channel attention)		92.8
CNN+Bi-GRU (spatial-attention)		93.6
DA-CNN+Bi-GRU		95.6
STAD-ConvBi-LSTM (Proposed)		**98.2**
CNN+LSTM	YouTube Action [154]	64.7
CNN+Bi-LSTM		84.2
CNN+GRU		88.5
CNN+Bi-GRU		92.1
CNN+Bi-GRU (channel attention)		94.2
CNN+Bi-GRU (spatial attention only)		95.6
DA-CNN+Bi-GRU		96.2
STAD-ConvBi-LSTM (Proposed)		**98.0**

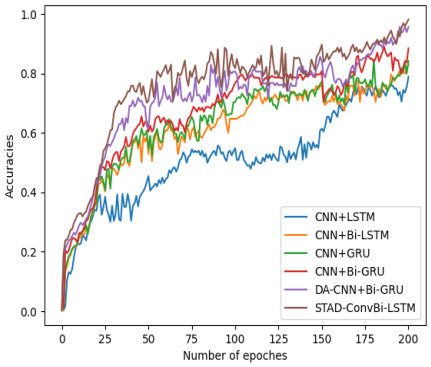
Approach	**Dataset**	**Acc %**
CNN+LSTM	HMDB51 [151]	56.7
CNN+Bi-LSTM		63.2
CNN+GRU		68.0
CNN+Bi-GRU		72.4
CNN+Bi-GRU (channel attention)		73.9
CNN+Bi-GRU (spatial attention only)		74.5
DA-CNN+Bi-GRU		79.3
STAD-ConvBi-LSTM (Proposed)		**80.7**
CNN+LSTM	UCF101 [149]	83.9
CNN+Bi-LSTM		86.8
CNN+GRU		90.7
CNN+Bi-GRU		94.2
CNN+Bi-GRU (channel attention)		95.1
CNN+Bi-GRU (spatial attention only)		95.8
DA-CNN+Bi-GRU		97.6
STAD-ConvBi-LSTM (Proposed)		**98.2**
CNN+LSTM	Kinetics-600 [155]	73.2
CNN+Bi-LSTM		77.9
CNN+GRU		81.5
CNN+Bi-GRU		84.5
CNN+Bi-GRU (channel attention)		90.2
CNN+Bi-GRU (spatial attention only)		92.6
DA-CNN+Bi-GRU		94.6
STAD-ConvBi-LSTM (Proposed)		**96.8**
CNN+LSTM	Synthesized Dataset	74.2
CNN+Bi-LSTM		82.9
CNN+GRU		85.0
CNN+Bi-GRU		92.7
CNN+Bi-GRU (channel attention)		92.8

Approach	Dataset	Acc %
CNN+Bi-GRU (spatial attention only)		94.7
DA-CNN+Bi-GRU		96.8
STAD-ConvBi-LSTM (Proposed)		**97.9**

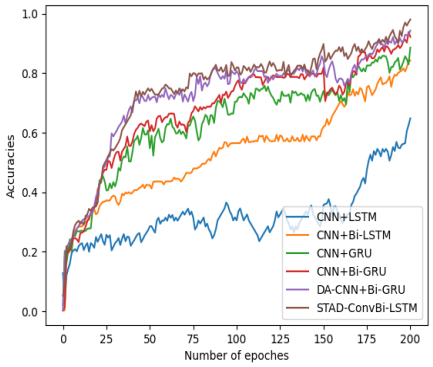
Fig. 4.6 depicts the training validation for each baseline method with our proposed framework. The graphical representation clearly shows the training accuracies of all baseline methods and offers the superior performance of STAD-ConvBi-LSTM. Additionally, Fig. 4.6a shows the comparison for the UCF50 dataset [153], where the STAD-ConvBi-LSTM method reaches the best accuracy over the 200th epochs. Similarly, in Fig. 4.6b for the YouTube Action dataset [154], the proposed method achieves a better accuracy score over the 200th epochs. Fig. 4.6c for the UCF101 dataset [149], our method performs the best in the first 30 epochs, where DA-CNN+Bi-GRU [157] leads and after that, the proposed methods again start rectifying and at the 200th epochs train out with the best accuracy. Fig. 4.6d for the HMDB51 dataset [151], our method starts with the best training accuracy in the very early epochs and shows the best performance throughout the training phase for 200 epochs, but DA-CNN+Bi-GRU leads at 50 epochs; meanwhile, after the dropout layer, it attains the best result. Fig. 4.6e for the Kinetics-600 dataset [155], our method does not perform the best in the first 25 epochs, where DA-CNN+Bi-GRU [157] leads and after that, it increases but again drops at the 60th epoch and after that rts rectifying over 200th epochs. Lastly, in Fig. 4.6f for the Synthesized dataset, the STAD-ConvBi-LSTM method attains the finest training accuracy all over the training phase of other methods. Table 4.3 illustrates the ablation studies of SOTA approaches over the proposed method on publicly available HAR benchmarks and observed that STAD-ConvBi-LSTM leads all baseline approaches over each dataset.

4.3.3.2 Comparison with existing SOTA methods

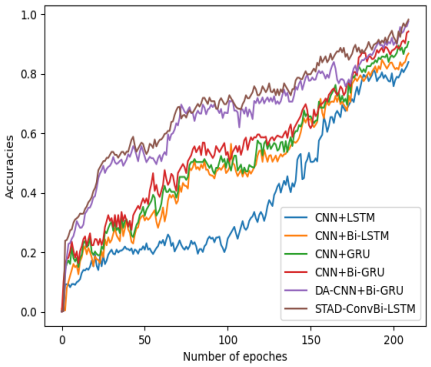
We conducted a comprehensive systematic analysis of the proposed STAD-ConvBi-LSTM framework with existing SOTA for AR approaches for evaluating overall accuracy and effectiveness. The quantitative comparisons for UCF50, YouTube action, UCF101, HMDB51, Kinetics-600 and our own synthesized datasets, respectively are represented in tables 4.4 to 4.9. The best and runner-up accuracy are highlighted in bold. Inspecting the results presented, it can be evident that the STAD-ConvBi-LSTM framework surpasses SOTA methods in performance on various datasets. Table 4.4 describes the accuracies for the UCF50 dataset [153]; STAD-ConvBi-LSTM



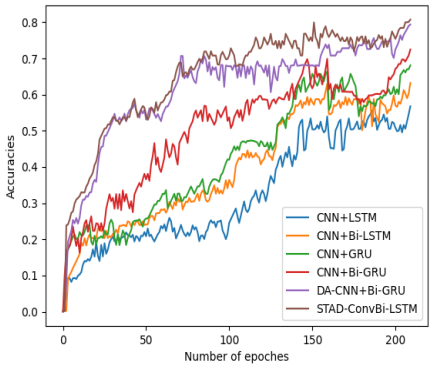
(a) Training Accuracies for UCF50 Dataset



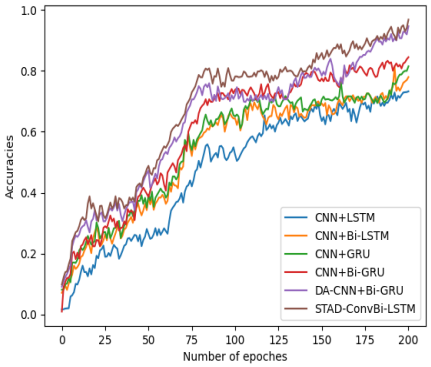
(b) Training Accuracies for YouTube Action



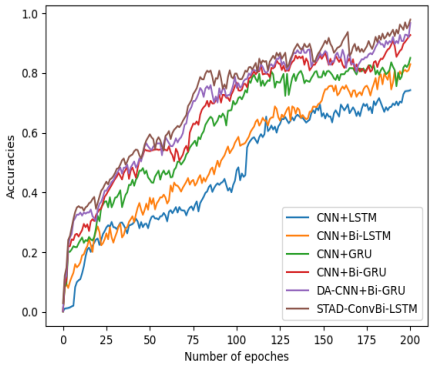
(c) Training Accuracies for UCF101 Dataset



(d) Training Accuracies for HMDB51 Dataset



(e) Training Accuracies for Kinetics-600 Dataset



(f) Training Accuracies for Synthesized Dataset

Fig. 4.6: Trade-off curves of proposed STAD-ConvBi-LSTM framework with other verified baseline methods over various HAR and synthesized dataset

framework leads the SOTA methods by attaining the best accuracy, 98.8%, while DA-CNN+Bi-GRU [157] obtains runner-up accuracy of 98.5%. The rest of the methods, including deep autoencoder [158], DS-GRU [59], Two-stream RGB DL-Net [60], and ViT+LSTM [159], obtained 96.4%, 92.2%, 95.4% and 96.1% accuracies, respectively. The hand-crafted method HOG [160] obtains very poor accuracy as of 24.8%, even not featuring the exact activity within the frames.

Table 4.4: Quantitative analysis of STAD-ConvBi-LSTM framework with SOTA Techniques on the UCF50 dataset

Techniques	**Accuracy (%)**
HOG [160]	24.8
Deep-Autoencoder [158]	96.4
DS-GRU [59]	92.2
Two-stream RGB DL-Net [60]	95.4
ViT+LSTM [159]	96.1
DA-CNN+Bi-GRU [157]	98.5
STAD-ConvBi-LSTM (Proposed)	**98.8**

Table 4.5 describes the accuracies for YouTube Action [154]. The ST-DAN [161] has the best performance, with an accuracy of 98.2%. STAD-ConvBi-LSTM attains the runner-up performance by obtaining an accuracy of 98.1% followed by DA-CNN+Bi-GRU [157] with have accuracy 98.0% Other methods compared include BI-LSTM [162], deep autoencoder [158], two-stream attention LSTM [163], dilated CNN+BiLSTM+RB [164], and Two-stream RGB DL-Net [60] which obtain 85.3%, 96.2%, 96.9%, 89.0% and 97.1% accuracies, respectively.

Table 4.5: Quantitative analysis of STAD-ConvBi-LSTM framework with SOTA techniques on YouTube Action dataset

Techniques	**Accuracy %**
Bi-LSTM [162]	85.3
Deep-Autoencoder [158]	96.2
ST-DAN [161]	98.2
Two-Stream Attention-LSTM [163]	96.9
Dilated CNN+BiLSTM+RB [164]	89.0
Two-stream RGB DL-Net [60]	97.1
DA-CNN+Bi-GRU [157]	98.0
STAD-ConvBi-LSTM (Proposed)	**98.1**

Table 4.6 describes the accuracies for HMDB51 dataset; STAD-ConvBi-LSTM obtains best accuracy of 81.2%, while the second-best method, Two-stream RGB DL-Net [60] attains an accuracy of 79.8%. The rest of the comparative methods like OF+multi LSTM [165], TSN [166], IP-LSTM [167], deep autoencoder [158], TS-LSTM+temporal-inception [168], correlational CNN+LSTM [169], ST-DAN [161], DB-LSTM+SSPF [170], ViT+LSTM [159], semi-supervised temporal gradient learning [171], & DA-CNN+Bi-GRU [157] obtain accuracies of 72.2%, 70.7%, 58.6%, 70.3%, 69.0%, 66.2%, 56.5%, 75.1%, 73.7%, 75.9%, 55.6% and 79.3%, respectively.

Table 4.6: Quantitative analysis of STAD-ConvBi-LSTM framework with SOTA approaches on the HMDB51 dataset

Techniques	**Accuracy %**
OF + multi-LSTM [165]	72.2
TSN [166]	70.7
IP-LSTM [167]	58.6
Deep-Autoencoder [158]	70.3
TS-LSTM+Temporal-Inception [168]	69.0
Correlational CNN+LSTM [169]	66.2
ST-DAN [161]	56.5
DB-LSTM+SSPF [170]	75.1
ViT + LSTM [159]	73.7
Semi-Supervised Temp-Gradient learning [171]	75.9
Two-stream RGB DL-Net [60]	79.8
DA-CNN+Bi-GRU [157]	79.3
STAD-ConvBi-LSTM (Proposed)	**81.2**

Table 4.7 describes the accuracies for UCF101 dataset [149]; the CNN+Bi-LSTM [172] method attains the best comparative result of 97.6% accuracy, followed by the second highest DA-CNN+Bi-GRU [157], have 97.5% accuracy. Our proposed method achieves the third-highest accuracy of 97.4%. The rest of the comparative methods including spatio-temporal multiplier networks [173], long-term temporal convolutions [174], attention cluster [175], Video-lstm [176], two-stream convnets [177], mixed 3D-2D convolutional tube [178], TS-LSTM+temporal-inception [168], correlational CNN+LSTM [169], Conv-Net Transformer [179], Two-stream RGB DL-Net [60] and HOG [160] achieve accuracies of 87.0%, 82.4%, 94.6%, 89.2%, 84.9%, 88.9%, 91.1%, 92.8%, 86.1%, 88.6%, and 29.2%, respectively.

Table 4.7: Quantitative analysis of STAD-ConvBi-LSTM framework with SOTA approaches on the UCF101 dataset

Techniques	**Accuracy %**
OG [160]	29.2
Spatio-Temporal Multiplier Net [173]	87.0
Long-Term Temporal Convolutional [174]	82.4
Attention-Cluster [175]	94.6
CNN+Bi-LSTM [172]	97.6
Video-lstm [176]	89.2
Two-stream ConvNets [177]	84.9
Mixed 3D-2D Convolutional Tube [178]	88.9
TS-LSTM+Temporal-Inception [168]	91.1
Correlational CNN + LSTM [169]	92.8
Conv-Net Transformer [179]	86.1
Two-stream RGB DL-Net [60]	88.6
DA-CNN+Bi-GRU [157]	97.5
STAD-ConvBi-LSTM (Proposed)	**97.4**

Table 4.8 describes the accuracies for the Kinetics-600 dataset [155]. The proposed method STAD-ConvBi-LSTM attains the finest performance of 88.2% accuracy, followed by DA-CNN+Bi-GRU [157] has an accuracy of 86.7%, while the Two-stream RGB DL-Net [60] achieves 3rd highest accuracy of 83.7%. The remaining comparative methods, such as Stnet [180], GCF-Network [181], and Global & Local-aware Attention [182], achieve accuracies of 76.3%, 70.0%, and 70.0%, respectively.

Table 4.8: Quantitative analysis of STAD-ConvBi-LSTM framework with SOTA approaches on the Kinetics-600 dataset

Techniques	**Accuracy %**
Stnet [180]	76.3
GCF-Network [181]	70.0
Global & Local-aware Atten. [182]	70.0
Two-stream RGB DL-Net [60]	83.7
DA-CNN+Bi-GRU [157]	86.7
STAD-ConvBi-LSTM (Proposed)	**88.2**

Table 4.9 describes the accuracies for the Synthesized Human Action Dataset; the STAD-ConvBi-LSTM attains the greatest accuracy of 96.7%, followed by STA-TSN

(RGB + Flow) [183] with an accuracy of 93.8%. The rest of the comparative methods including Spatio-Temporal Multiplier Networks [173], Long-Term Temporal Convolutional [174], CNN+Bi-LSTM [172], Correlational CNN + LSTM [169], Global and local-aware attention [182], TS-LSTM + Temporal-Inception [168], Two-stream RGB DL-Net [60] and DA-CNN+Bi-GRU [157] achieve accuracies of 74.3%, 76.2%, 80.6%, 83.4%, 72.1%, 83.6%, 83.7%, and 86.7%, respectively.

Table 4.9: Quantitative analysis of STAD-ConvBi-LSTM framework with SOTA approaches on the Synthesized Human Action Dataset

Techniques	**Accuracy** %
Spatio-Temporal Multiplier Networks [173]	74.3
Long-Term Temporal Convolutional [174]	76.2
CNN+Bi-LSTM [172]	80.6
Correlational CNN + LSTM [169]	83.4
Global & Local-Aware Atten. [182]	72.1
TS-LSTM+Temporal-Inception [168]	83.6
Two-stream RGB DL-Net [60]	83.7
STA-TSN (RGB + Flow) [183]	93.8
DA-CNN+Bi-GRU [157]	86.7
STAD-ConvBi-LSTM (Proposed)	**96.7**

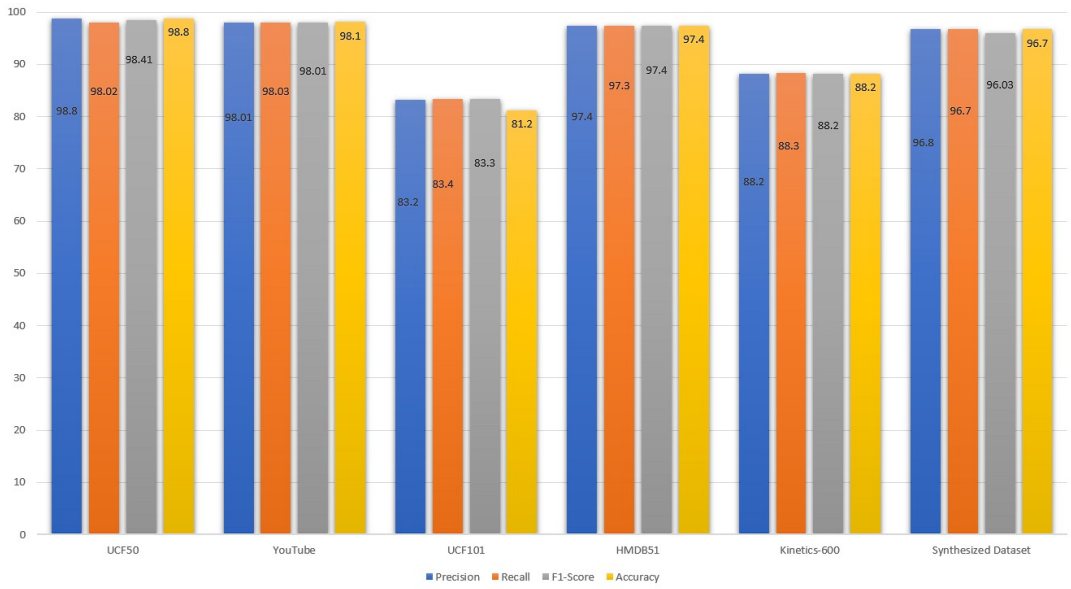


Fig. 4.7: Relative Stats of Performance Metrics on various Human Action Recognition datasets

The performance of the model has been measured on the different performance metrics i.e., Precision, Recall, F1-score, and Accuracy. Fig. 4.7 represents the accuracy, precision, recall, and F1-score scores on different benchmarks. While considering the overall comparative analysis, our STAD-ConvBi-LSTM offers the best trade-off between computational complexity and action recognition accuracy and greatly leads the comprehensive SOTA methods. Fig. 4.8 shows the representative frames of the predicted activity clips along with their ground truths, model predicted class, and accuracies and observed that the performance of the model is sufficiently high.

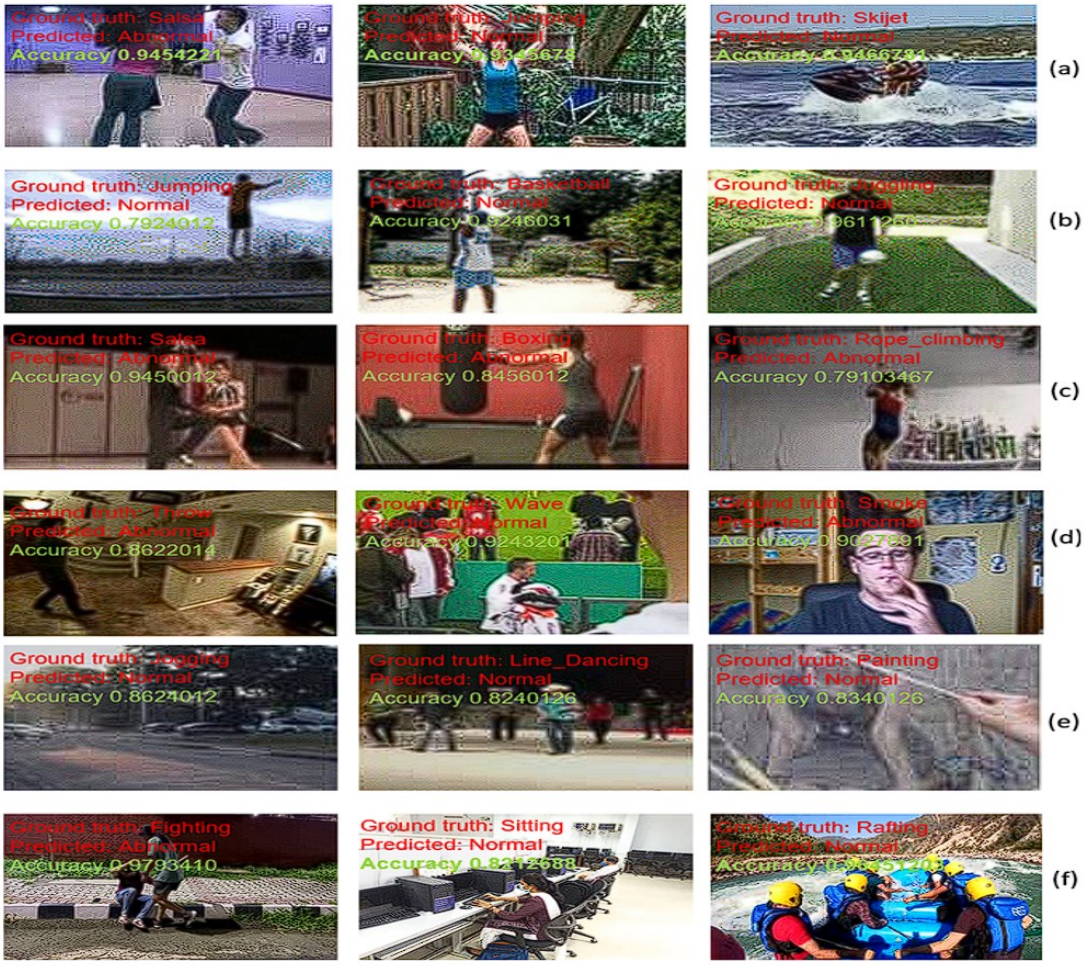


Fig. 4.8: Illustration of Human Action Recognition on various benchmarks. Starting from top–bottom (a) UC50 Dataset (b) YouTube Action Dataset (c) UCF101 Dataset (d) HMDB51 Dataset (e) Kinetics-600 Dataset (f) Synthesized Dataset.

4.4 Summary

This chapter presented a channel-wise spatial-temporal discriminative AR learning framework in video clips. The proposed framework captures the spatial attentional CNN architecture with a Bi-LSTM-net for solitary occurrence training and extraction of efficient spatial-temporal key features of human activities. Firstly, we employ CNN architecture, which contains spatial attention for extracting the relevant features from video, specifically human activity-specified regions, and producing high-level feature representation. The bi-LSTM also learns the temporal modelling of long-term human activity sequences through a bi-directional passway (i.e., forward & backward) and helps the proposed framework to learn information from the preceding frames as well as from the succeeding sequence frames. This bi-directional modelling for human activity significantly enhances our approach to learning capability during training phase and improves prediction precision. For the estimation of the STAD-ConvBi-LSTM framework efficiency calculated on various publicly available benchmark HAR datasets. The attained experimental results achieve better performance as correlated with existing SOTA methods & verify the effectiveness with respect to models' robustness and computational efficiency.

Chapter 5

Occluded Skeleton-Based Multi-Stream Model using Part-Aware Spatial-Temporal Graph Convolutional Network for Human Activity Recognition

In this chapter, we introduce a novel occlusion-resilient framework for skeleton-based human activity recognition, termed as MSPAST-GCN. The proposed approach integrates a multi-stream part-aware spatial-temporal graph convolutional network, designed to effectively extract both spatial and temporal features from occluded skeletal input. By employing an inhibition strategy, the model enhances its focus on informative, uninhibited keypoints, leading to improved spatial feature learning. The framework also introduces PAST-GCN, which leverages part-aware spatial and temporal graph convolutions to boost representation learning.

5.1 Introduction

Skeleton-based in HAR approaches provide a more accurate structure and relevant information than other modalities like RGB, Optical Flow, and deep learning-based GCN. It is highly robust to variations in illumination, intensity, and different backgrounds, combined with its low-dimensional feature representation, significantly conserving computing resources. Due to progress in depth sensor technology, skeleton data have become increasingly accurate and accessible. These improvements have contributed considerably to the growing status of skeleton-based HAR and become a preferred choice in the research areas of computer vision [184]. Since skeleton input sequence can be regarded as a graph data type, it offers a more accurate and comprehensive representation, such as graph structure in the form of edges and body key-joints. GCN-based methods have become progressively popular and attained significant achievement [185, 186, 187, 130, 61]. However, these methods often fail

while addressing common problems such as occlusion and multi-interaction. When an object obstructs critical body key-joints, the recognition capabilities of these models are substantially compromised. Despite the rapid advancements in skeleton-based HAR techniques, improving the robustness of models, especially in occluded environments, remains a significant challenge. Several research studies have adeptly designed GCN to address the problem based on noisy data or incomplete skeletons to improve the model's effectiveness and increase efficiency. However, these approaches fail under different illumination conditions. In realistic scenarios, capturing incomplete or noisy skeletons from the given input is almost inevitable, and more robust solutions are needed. For instance, translucent or opaque objects may hide the targeted human activity, and their prediction of the activity can be affected by some environmental aspects, such as background conditions and fluctuating illumination. Previously, the traditional models failed to acknowledge this part and focused only on extracting discriminative body key-joints. In the case of occluded body key-joints, the models fail to predict the correct activity. These limitations highlight the need for models capable of maintaining high performance, even when body key-joints are not visible.

Thus, to solve such problems, introducing a multi-stream part-aware occluded skeleton spatial-temporal GCN model termed as MSPAST-GCN, as depicted in Fig. 5.1, that works on occlusion to drive the model for activating the body key-joints related activity and extract the relevant features. Within the whole network, we first inhibit body key-joints in the input sequence to pretend scenarios where critical portions are occluded using the inhibition training strategy. Conversely, we divide the predicted score map within grids and then arbitrarily inhibit specific values in the grids. The inhibition strategy not only pretends arbitrary occlusion but also raises the prediction challenge, compelling the model to acquire and extract more prominent features like key-joint coordinates, relative key-joint coordinates, and temporal differences from the uninhibited joints. This strategy improves the model's robustness on incomplete skeleton data. The proposed model improves the recognition efficiency of activity on occluded skeleton sequence data via pretending occlusion conditions and extracting the spatial-temporal feature. The model entails three modules: Input Inhibition Module for Skeleton Sequences (IIMS), Part-Aware Spatial-Temporal Graph Convolutional Network (PAST-GCN), and Predicted Score Inhibition (PSI). The IIMS module pretends occlusion by inhibiting decisive key-joints in the input skeleton data, preparing the model to handle real-world occlusion scenarios. The PAST-GCN module extracts local and global spatial-temporal discriminative features from skeletons, enhancing the model's capability to learn associations between key-joints and precisely enhancing robustness. The PSI module improves the prediction difficulty by randomly inhibit-

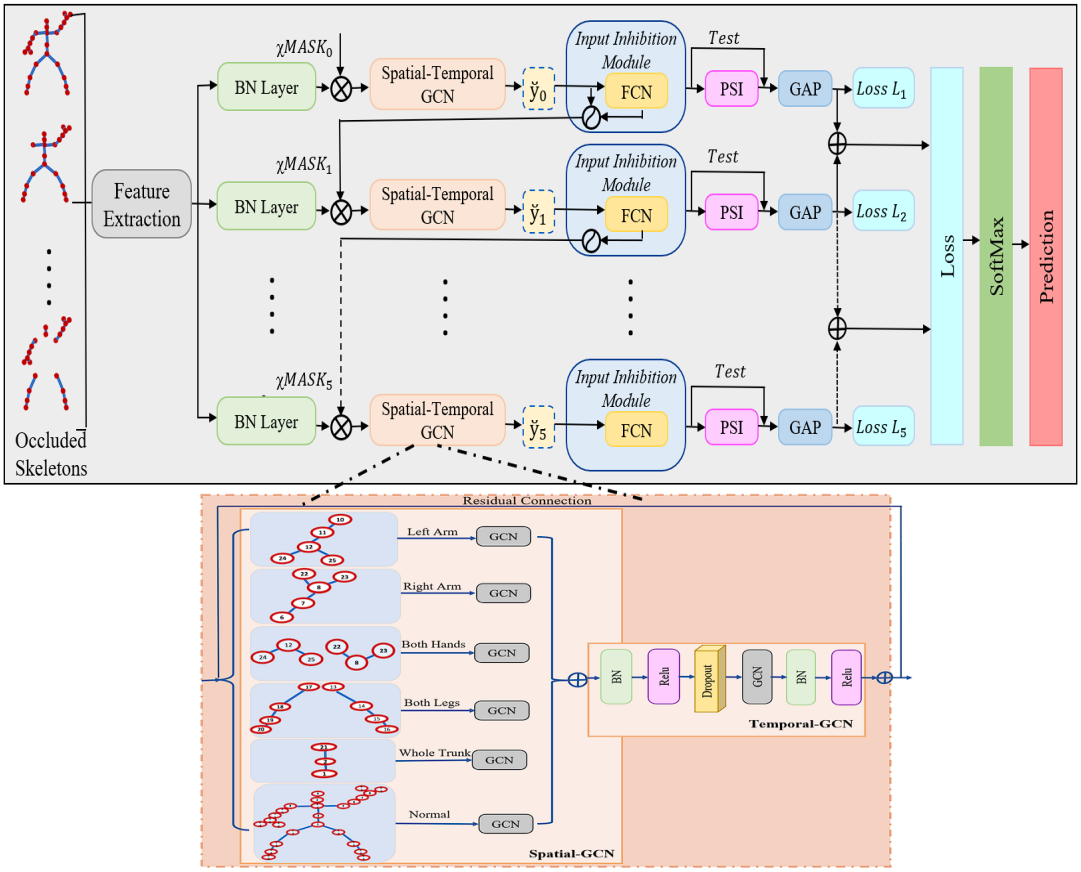


Fig. 5.1: Pipeline model of the proposed MSPAST-GCN. It's a multi-stream model that extracts the spatial-temporal features and predicts the class of an activity, where each stream contains three modules: PAST-GCN, IIMS, and PSI.

ing parts of the predicted score map, driving the model to extract more discriminative features from the uninhibited key points. Furthermore, built a synthesized occluded dataset comprising numerous occlusion conditions and designed several experimental settings to rigorously evaluate model performance from different observations. The proposed model performs better with SOTA methods on three synthesized occluded datasets. The key contributions are as follows:

- We propose Occluded Skeleton-Based Multi-Stream Part-Aware Spatial-Temporal GCN for human activity recognition within skeleton input sequences.
- The inhibition strategy is utilized to advance the model's learning, thereby improving the ability to extract the spatial dimension features among uninhibited key points.
- Subsequently, the Part-Aware Spatial-Temporal GCN termed as PAST-GCN is introduced to extract information on spatial key-joint features. Simultaneously,

the model integrates a temporal graph convolution to improve the ability to extract the temporal dimension features.

- The proposed MSPAST-GCN model outperforms SOTA performance on benchmarks, i.e., NTU RGB+D 60, NTU RGB+D 120, and RGBD-Action-Completion-2016. Also, constructed a synthesized dataset that comprises various occlusion conditions.

5.2 Proposed Methodology

Occlusions can reduce the effectiveness of HAR in real-time scenarios. Human bodies may be partially or fully occluded in video datasets or images, which decreases activity recognition accuracy. Two categories of occlusion are there i.e., spatial and temporal occlusion. Spatial occlusion refers to the partial obstruction of key joints in individual frames due to external factors like furniture, objects, walls, etc. Temporal occlusion is defined as missing sequences in the temporal domain, where frames (video frames for particular time series)are partially or entirely absent. Most methods focus on optimizing performance for non-occluded data, often ignoring the model's ability to remain robust in occlusions. Thus, to tackle this issue, we introduced the multi-stream part-aware spatial-temporal graph convolution network using an inhibition strategy called MSPAST-GCN. The working channel of our proposed MSPAST-GCN is demonstrated in Fig. 5.1. First, the input skeleton sequences extract features and mask the diverse local and global parts of Spatial-Temporal Graph Convolutional (ST-GCN) subnetworks. Conversely, we simulate occlusion by employing an inhibition strategy using a predicted confidence inhibition module. Each ST-GCN subnetwork has a diverse input mask for distinguishing the six occlusion cases and identifying the group of the activity classes. Lastly, the output of all streams has been integrated to attain the final prediction. While simultaneously employing the cross-entropy to compel the final result prediction and the predictions of each stream, enabling all streams to be optimized together.

The proposed MSPAST-GCN model introduces an inhibition strategy that simulates occlusions during training, forcing the model to rely on alternative, uninhibited key joints for classification. This strategy operates at two levels: (i) input inhibition and (ii) predicted score inhibition. Input inhibition selectively masks key-joint coordinates in skeleton data to mimic real-world occlusions, enhancing the model's ability to recognize activities with missing information. Predicted score inhibition modifies confidence scores by reducing emphasis on certain regions in the feature map, preventing

over-reliance on visible joints, and encouraging better generalization. It is important to highlight that we don't inhibit the predicted score map throughout the testing phase. Additionally, the MSPAST-GCN method supports multi-stream configurations to comprehensively capture the spatiotemporal dynamics of human activities, i.e., 2-stream and 3-stream MSPAST-GCN.

5.2.1 Problem Statement

HAR using skeleton data plays a crucial role in surveillance and behavior analysis. However, occlusions caused by environmental objects, camera angles, or self-occlusion significantly degrade recognition accuracy. Existing GCN-based models face problems in handling occluded sequences effectively, as missing keyjoint information results in suboptimal feature representation.

To formally define the problem, let: $S = \{X_t\}_{t=1}^{T}$ be a sequence of skeleton frames, where each frame $X_t \in \mathbb{R}^{N \times D}$ contains N body keyjoints with D spatial dimensions (e.g., $D = 3$ for 3D coordinates). The goal is to design a model f_{out} that maps occluded skeleton sequences to an accurate activity label $\hat{y}$, ensuring robustness to missing information:

$$\hat{y} = f_{\text{out}}(X'_t) \tag{5.1}$$

where $\hat{y}$ is the predicted activity label. The model must compensate for missing keyjoint information by leveraging spatial-temporal dependencies & generalize across different occlusion types and durations without requiring explicit reconstruction of missing keyjoints.

5.2.2 Feature Extraction for Input Inhibition Skeleton Sequences Module

This module simulates occluded skeleton sequences directly during training to learn feature vector representations from different body keyjoints. The feature vector includes geometric feature representations, i.e., coordinates and relative keyjoint coordinates. A total of 25 keyjoints are taken as input sequences. The keyjoint coordinates χ are represented as:

$$\chi = \{x_{tij} \mid t = 1, \ldots, T,\ i = 1, \ldots, N\} \tag{5.2}$$

where x_{tij} is the keyjoint coordinate of i^{th} node with keyjoint j^{th} at t^{th} frame time. T indicates the input sequence length for the number of keyjoint nodes N. The relative

coordinates χ_r are represented as:

$$\chi_r = \{x_{tij} - x_{tcj} \mid t = 1, \ldots, T,\ i = 1, \ldots, N\} \tag{5.3}$$

where x_{tij} is the key node and x_{tcj} is the center node for key joint j. We concatenate key joint coordinates for the actual skeleton sequence of equations (5.1)-(5.2) for feature vector representation. The extracted feature vectors are examined as follows:

$$\chi_m = \text{concat}(\chi, \chi_r) \tag{5.4}$$

Then, we mask the random key joints of the input skeleton at diverse times to simulate real-world occlusion using the following equation:

$$\tilde{\chi} = \chi \otimes \text{MASK}_i \tag{5.5}$$

where MASK_i ($i = 0, 1, 2, \ldots, n$) represents the binary mask matrix for $t \times j$. $\otimes$ denotes element-wise multiplication. Taking inspiration from Grad-CAM approaches [188, 189], we mask each frame's videos of class activation maps for feature representation of human key joints. Suppose $y(t, j)$ represents the feature representation vector for the j^{th} key joint and t^{th} frame. Thus, the activation matrix A is calculated as:

$$y(t, j) = \sum_{c}^{i} W_C y(t, j) \tag{5.6}$$

where $y(t, j)$ represents elements of activation matrix A for the j^{th} key joint and t^{th} time, and W_C depicts the weight for the c^{th} class of activity. SoftMax is used to normalize the mask range between 0 and 1. The normalization mask $\mathcal{M}_k$ has a similar dimensionality as matrix A. The output value of $\mathcal{M}_k$ embodies the degree of input inhibiting and is calculated as:

$$\mathcal{M}_k = 1 - \text{SoftMax}(A) \begin{cases} \text{low}, & \mathcal{M}_k < 0 \quad \text{(unmasked joints)} \\ \text{high}, & \mathcal{M}_k \geq 0 \quad \text{(completely masked joints)} \end{cases} \tag{5.7}$$

To attain maximum consistency, we use a random generation technique to create MASK_0 for the first subnet. Additionally, to prevent noisiness with key-joint coordinates in the consequent subnet stream, we eliminate MASK_0 from the addition of the mask matrix in other streams. To maintain the model's robust performance despite significant information loss, we feed only those key joints that remain inactivated by the

second or third stream, which helps the model extract and learn effective geographical discriminative feature information from these unmasked key joints.

5.2.3 Part-Aware Spatial-Temporal Graph Convolution Module

In most scenarios, human activity does not involve the whole body; instead, it focuses on the action performed by specific body parts. For instance, taking glasses, eating, and throwing are under the categories of action performed via hands or arm key joints. Based on each parts of key joint nodes (1-25 joints), we construct a topological fully connected graph to highlight the localized features of HAR, as shown in Fig. 5.2. The skeleton structure is embodied as a sequence of vectors within each frame in which each vector shows 2D or 3D key joint coordinates.

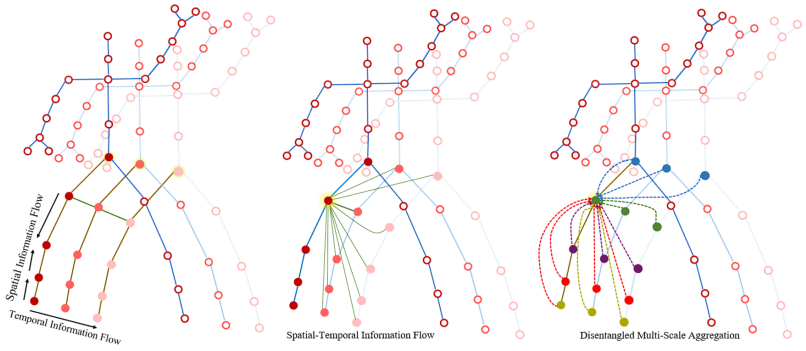


Fig. 5.2: Topological Fully Connected Graph Construction based on 25 key joints in the human body.

Earlier approaches [190] have pre-defined graph connections of key joints similar to ST-GCN for exhibiting local features representation with higher precision. This method partitions the keyjoint based on distance within adjacent keyjoints and center keyjoint, thereby simulating local features. These pre-defined connections and partitions ensure that the model focuses on relevant local relationships, enhancing its ability to discriminate the variations in motion and posture as per the keyjoint coordinates. We propose a Part-Aware Spatial-Temporal GCN Module consisting of Spatial-GCN (SGCN) and Temporal-GCN (TGCN) with a residual connection. The SGCN subnet is developed to show spatial features in detail. Based on joint semantic information, we partition the input data into different parts and create a matching topology graph by considering each key joint's part as fully connected. TGCN is similar to ST-GCN.

To extract the local feature representation, each part χ_{m_i} arrives at a different subnet. To obtain the global feature representation $\tilde{y}_0$, we firstly pre-define a graph on

skeleton keyjoint coordinates, then we feed the whole sequence χ_m to GCN. The resultant of each subnet GCN $\tilde{y}_k$ is determined by:

$$\tilde{y}_k = \sigma(D^{-\frac{1}{2}} A_k D^{-\frac{1}{2}} \chi_k W_k) \tag{5.8}$$

where k indicates the number of occluded joint parts with key joint coordinates j of the i^{th} ith node. χ_k is the occluded key joint, where χ_m represents the whole key joint. D specifies the degree matrix. We classify the relation among intra-frame key-joints within the similar coordinates of key joints in consecutive frames. Let's suppose A indicates the adjacency matrix for the i^{th} node and j key-joint coordinates. W_k indicates the weight matrix learned by the downstream subnet. The overall output of subnet GCN $\tilde{y}_k$ can be estimated as:

$$\tilde{y}_{k_{ij}} = \alpha \sum_{k-1}^{m} \tilde{y}_k + (1 - \alpha)\tilde{y}_0 \tag{5.9}$$

where α is the hyperparameter, and m indicates the number of masked occluded key-joints. This technique lets the model learn both global features of skeletons comprehensively. If the i^{th} and j^{th} key-joint coordinates are connected, then $A_{ij} = 1$ or $A_{ij} = 0$. For spatial feature representation, we apply spatial graph convolution, i.e., 'Convs', which can be formulated as:

$$f_{\text{out}}(v_i) = \sum_{v_j \in B_i} \frac{1}{Z_{ij}} f_{\text{in}}(v_j)\omega(l_i(v_j)) \tag{5.10}$$

where B_i, $f_{\text{in}}(v_j)$ signify the set of adjacent coordinates of key-joints v_i, and the feature information, respectively. $l_i(v_j)$ indicates the index of the adjacent coordinates (0, 1, 2) taken from [191]. Z_{ij} acts as a normalization term. ω indicates the weight equivalent to the adjacent coordinates of $l_i(v_j)$. For graphical structured image data, we map the value of each key-joint coordinate of the convolution kernel directly to the graphical image pixel extracted from video frames as skeleton. In contrast, we need to organize the corresponding adjacent coordinates of each key-joint for unstructured skeleton data. For temporal feature representation, we apply temporal graph convolution 'Convt'. In this, only the exact key-joint coordinates in corresponding adjacent frames are associated. Thus, we execute Convt using a $K_t \times 1$ convolution kernel of size, where K_t signifies the total number of adjacent video frames in skeleton graph. This approach facilitates the fusion of information across K_t adjacent frames.

To leverage spatial-temporal features more efficiently, we consecutively apply Convs and Convt, complied via batch normalization (BN) followed by a ReLU layer, respec-

tively. Subsequently, we incorporate a dropout layer among spatial-temporal convolutions to randomly drop some features and avoid the model from overfitting. Moreover, leveraging the principle that residual learning can speed up the network convergence [192], we add a skip connection between the input and output feed of the complete ST-GCN module, as shown in Fig. 5.3.

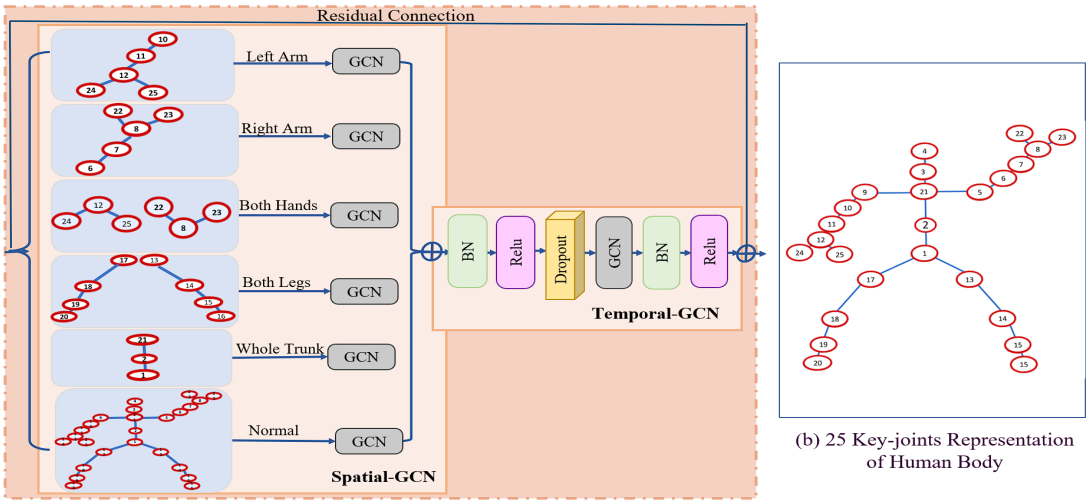


Fig. 5.3: (a) Construction of a ST-GCN module consist of Spatial-GCN and Temporal-GCN. (b) Represents the human body 25 key-joints

5.2.4 Predicated Score Inhibition Module

To handle the randomly occurring occlusion with indefinite time or location, we formulate a predicated score input employing random inhibit predicted scores map in Part-Aware ST-GCN dimensions of the skeleton sequence data as epitomized in Fig. 5.3. The modification block of fine-grained image recognition [193, 194] seeks to identify variations among related activity or object classes, which serves as a motivation for this module termed Predicted Score Inhibition (PSI) as depicted in Fig. 5.4. Further, this module is segregated into two parts, i.e., peak input inhibition and patch input inhibition.

5.2.4.1 Peak Input Inhibition

Let's assume the resultant predicted inhibited score map P_m of the fully connected layer is calculated as follows:

$$P_m(t, j) \in \tilde{y}^{C \times T \times J} \tag{5.11}$$

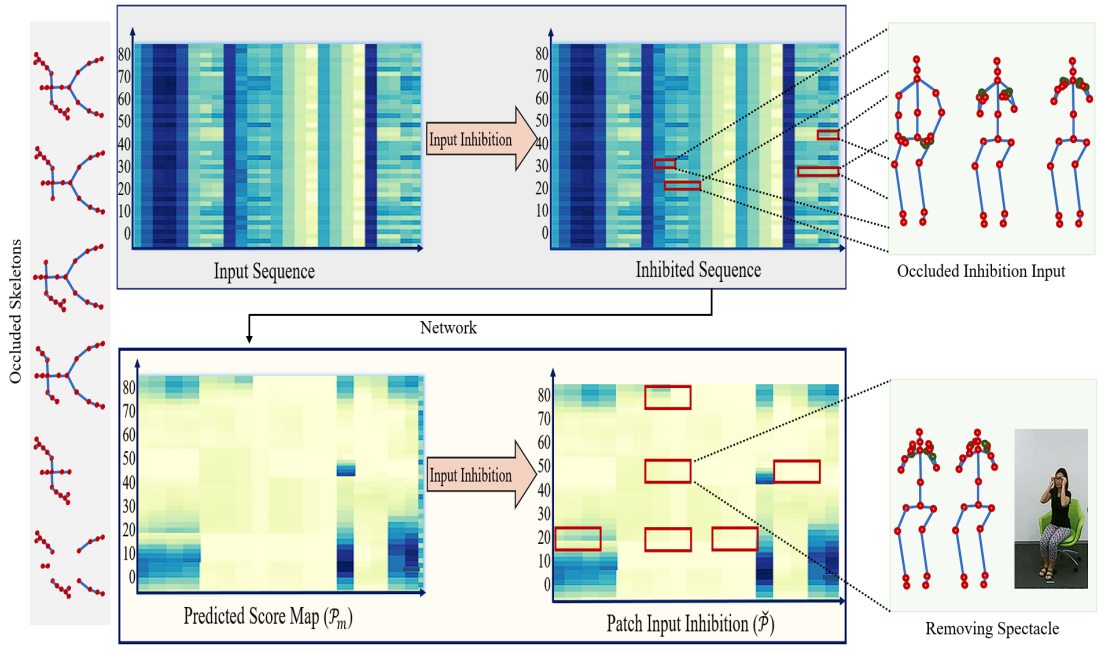


Fig. 5.4: Working of PSI Module consists of Peak Input Inhibition and Patch Input Inhibition. The sea-green rectangular area embodies the inhibited input area. The right side displays a visualization of the human skeleton, highlighting a specific area of inhibition where the red dot indicates the corresponding coordinates, while the green dot represents the inhibited score for the occluded portions.

where C indicates the activity classes and T indicates the input sequence length with several key-joint coordinates, such as N. This allows us to inhibit the key-joints by locating the peak value position of P_m, which is the prominent area for the classifier. We then employ a probability of P_{peak} to inhibit it. Then, the binary peak inhibition matrix of P_m is defined as:

$$P'(t,j) \in \tilde{y}^{T\times J} \quad , \quad P'(t,j) = \begin{cases} P_{\text{peak}}, & \text{if } P_m(t,j) = \max(P_m) \sim P_{\text{peak}} \text{ (Bernoulli)} \\ 0, & \text{otherwise} \end{cases} \tag{5.12}$$

The $\max(P_m)$ indicates the maximum function, and P_{peak} (Bernoulli) represents the Bernoulli random variable with the probability of P_{peak}.

5.2.4.2 Patch Input Inhibition

The use of patch input inhibition for the cases where certain human body parts are occluded in multiple consecutive frames to identify the other prominent areas. Further, we divide the predicted score map P_m into grids through spatial-temporal dimensions inspired by [195]. We use the probability of P_{patch} to inhibit each patch for calculating

the output patch as:

$$\text{patch}^{[l,n]} \in \tilde{y}^{T\times J} \quad \begin{cases} l \in [1,\ldots,T/t'], \\ n \in [1,\ldots,J/j']. \end{cases} \tag{5.13}$$

where t' and j' indicate the skeleton sequence length and total number of key-joints in the patch$^{[l,n]}$, respectively. Suppose $P''(t,j) \in \tilde{y}^{l'\times n'}$ represents the binary arbitrary patch inhibition matrix and is associated with a patch. If set to 1, the patch is inhibited, and set to 0 otherwise. Therefore, the P'' is evaluated as:

$$P'' = \begin{cases} \text{patch}^{(l,m)}, & \text{if patch = 1} \\ 0, & \text{otherwise} \end{cases} \quad \in \tilde{y}^{l'\times n'} \quad \begin{cases} l = [1,\ldots,T/t'], \\ m = [1,\ldots,J/j']. \end{cases} \tag{5.14}$$

The total inhibition matrix $P \in \tilde{y}^{L\times N}$ is estimated by the aggregation of the peak inhibition matrix and patch inhibition matrix, and is thus formulated as:

$$P = 1 - \beta \cdot (P' \mid P''). \tag{5.15}$$

where β acts as an inhibition factor and $\mid$ signifies the operator. In conclusion, the resultant inhibited score map is estimated as:

$$\hat{P} = P \odot P_m \tag{5.16}$$

By applying the techniques of patch inhibition and peak inhibition, we increase the model's learning capability and improve the achievement of discriminative features from uninhibited key-joints, strengthening the model's resilience. The entire predicted score map P_m is fed to the GAP layer during the test phase without any inhibition in any area. Thus, every information region identified throughout the training phase adds to the total confidence score.

5.2.5 Multi-Stream Graph Convolutional Networks (2-Stream and 3-Stream Variants)

The proposed MSPAST-GCN method supports multi-stream configurations to comprehensively capture the spatiotemporal dynamics of human activities, specifically 2-stream MSPAST-GCN and 3-stream MSPAST-GCN. The 2s MSPAST-GCN utilizes two input streams: Key Joint Stream & Bone Stream. Key Joint stream captures

spatial relationships and temporal dependencies among body key joints. While bone stream captures the complementary motion information by analyzing the dynamics of skeleton bones. These two streams are processed independently and fused at the final classification stage to enhance overall performance. Extending this, the 3s MSPAST-GCN introduces Motion Stream also, which explicitly encodes motion patterns over time by capturing temporal differences between consecutive frames. This three-stream framework leverages complementary information from key joints, bones, and motion, enabling the model to capture fine-grained spatial and temporal features.

5.2.6 Loss Optimization Strategy

We employ cross-entropy as a loss function to recognize occluded activities and put constraints during training on each stream of the network because each stream attempts to appropriately predict the activity class. The cross-entropy constraint is formulated via:

$$\text{Loss-Function}_i = -y \log \tilde{y}_i \tag{5.17}$$

where Loss-Function$_i$ is the loss constraint for the i^{th} stream, y represents the ground-truth activity label, and $\hat{y}_i$ is the predicted probability from the i^{th} stream.

The final prediction's cross-entropy constraint is:

$$\text{Loss-Function}_p = -y \log \tilde{y}_i \tag{5.18}$$

The loss function for key-joint of sub-streams during training is given as:

$$\text{Loss-Function}_j = -y \log \tilde{y}. \sum_{i=0}^{6} y \log \tilde{y}_i \tag{5.19}$$

where summation $\sum_{i=0}^{6}$ indicates that the loss is computed independently for each of the six occluded parts, and the total loss is aggregated across these parts. Initially, we defined six categories of occlusion masks corresponding to different scenarios: no occlusion (NO), two legs (TL), two hands (TH), left arm (LA), right arm (RA), and whole trunk (WT) occlusions, as shown in Fig. 5.5. The network is trained to extract both local and global features by adaptively detecting occlusion and enhancing activity recognition accuracy to mitigate the effects of occlusion. During training, we randomly inhibit partially occluded skeletal data and select one of these occlusion cases to mask the multimodal features. During testing, we directly evaluate the performance of various approaches for activity recognition using the occluded dataset.

5.3 Experimental Analysis

This section describes unoccluded and occluded datasets with the experimental settings. It also discusses the experimental result analysis. We performed an ablation review to validate each proposed component's effectiveness.

5.3.1 Benchmark

5.3.1.1 Unoccluded Dataset

NTU-RGB+D 60 Dataset [113] is presently the utmost extensively used indoor motion activity recognition due to the largest amount of data and the most abundant data modalities. It comprises 56,880 trial samples recorded via Microsoft Kinect v2, and each sample consists of 4 diverse modalities. We take the 3D skeleton sequence for activity recognition, consisting of x-y-z coordinates with 25 human body key-joints. For 60 categories of activity classes recorded via 40 volunteers at 10-35 ages, and each activity was recorded via three cameras positioned at equal height with diverse angles. The benchmarks are divided into Cross-Subject subset (CS) and Cross View subset (CV). Out of 40 subjects, the CS subset includes a training sample of 40,320 sets and testing samples of 16,560 sets, whereas the CV subset includes a set of 37,920 training sets along the camera's 2nd & 3rd, and 18,960 testing samples recorded via 1st camera.

NTU-RGB+D 120 Dataset [89] is the lengthy form of above dataset, including a supplementary of 57,367 samples covering 60 different classes of human activity. Further, the benchmark is divided into two test subsets, namely Cross-Set (CSET) and Cross-Subject (CSUB). In CSET, 54,468 activity samples were taken for training, and 59,477 were taken for the testing set, whereas CSUB included a training set of 63,026 and 50,919 testing samples, respectively.

RGBD-Action-Completion-2016 Dataset [193] was released at the University of Bristol and comprised 414 sample videos of some complete and some incomplete activity captured via a Microsoft Kinect v2. The frames include three categories: RGB, mask-depth, and skeleton sequence. We choose skeleton sequence only in which each skeleton has 25 human body key-joints captured on the x-y-z axis coordinates. Overall, six activities are recorded to indicate human and object interactions, i.e., switching the light off-on, plug-in-socket, opening-jar, pulling-drawer, pick-up an object from the desk, and drinking from a cup.

5.3.1.2 Occluded Synthesized Dataset

Occlusion in any human activity sequence of video is diverse and highly unpredictable. The extent of the occluded portion, its size, and its occurrence timing are all uncertain factors. Thus, activity recognition under these conditions offers significant challenges. As per, our studies, no public HAR datasets are available or designed to address partial occlusion scenarios. This gap presents an important challenge for researchers looking to test and improve their methods under such conditions. To address this, we constructed a synthetic dataset containing diverse occlusion scenarios to simulate realistic conditions. This dataset includes 20% of the samples from the CSET view of NTU RGB+D 60, 20% of the samples from the CSET view of NTU RGB+D 120, 20% of the samples from RGBD-Action-Completion-2016, and the remaining 40% from our own dataset, which is developing in DTU campus and its surrounding areas. So far, out of 350 planned video clips, 300 have been synthesized. The video clips were recorded using a 12-megapixel camera with f/1.6 aperture, 26 mm wide lens, 1/1.9" sensor, 1.7 μm pixel size, PDAF dual-pixel autofocus, and sensor-shift optical image stabilization (OIS) at a resolution of 480*640. The dataset currently consists of six activity classes: Adverse, Sports, Vandalism, Normal, Shoplifting, and Occlusion categories. It includes 250 video clips, each lasting 5–6 seconds, with 60 video clips per class. We randomly chose 60% of the data sequences from them. Our occlusion experimental setup includes six types of occlusions (i.e., Two-hand (TH), Two-Leg (TL), Right-Arm (RA), Left-Arm (LA), Whole-Trunk (WT), and one with no occlusion case (NO)). These six cases for training and testing are illustrated in Fig. 5.5.

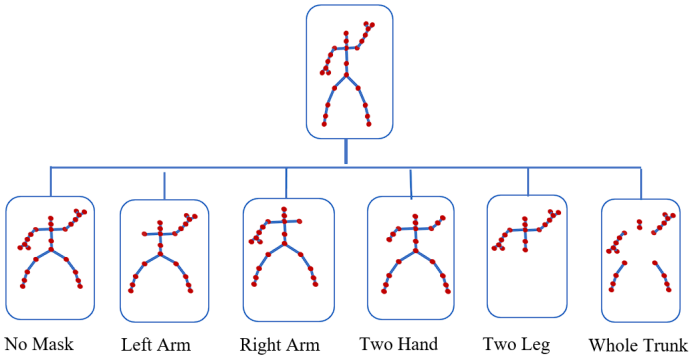


Fig. 5.5: Six Test Occluded Cases in Experimental Setup.

5.3.2 Implementation Setup Details

5.3.2.1 Network Details

The experiments are conducted on GPU NVIDIA RTX 4060 64GB RAM AMD Ryzen 9 in Python using PyTorch.During the training phase, Stochastic Gradient Descent is utilized as the optimization algorithm, with a momentum of 0.9 to stabilize training. The model is trained with a cross-entropy loss function, using a batch size of 16, a maximum of 200 epochs, and an initial learning rate of 0.1. To mitigate the risk of overfitting, dropout with a probability of 0.5 is applied at each GCN stream. A BN layer is applied at the input stage of the network to normalize the data. In SGCN (Spatial-based Graph Convolutional Network) module within the ST-GCN module, the human body is divided into 5 non-overlapping parts: TL, TH, LA, RA and WT. These body parts are treated as interconnected to capture potential associations and relationships among key joints as determined in Eq.5.7,5.8. The proportional hyperparameter α is set to 0.1, determining the strength of interaction between different body parts. In IIM module, the number of neurons in FC layer is set to match the number of test cases, corresponding to the classification task. The PSI module introduces hyperparameters such as inhibition probability, patch width, patch length and patch size and adjusts based on the dataset's characteristics. The inhibition probability set as P_{peak} ($\mathcal{P}'$)=0.9, P_{patch} ($\mathcal{P}''$)=0.2, and occlusion factor β=0.1 utilized to control the proportion of inhibition applied across different body parts, preventing extreme occlusions that might degrade model learning stability as in Eq.5.11. Patch width j'=5, and patch length t'=10 as in Eq.5.12, respectively. Lastly, an average GAP is applied to the predicted inhibited score map with spatial-temporal dimensions to produce the prediction score for each sub-stream.

5.3.2.2 Experimental Setting

To comprehensively evaluate occluded skeleton-based activity recognition from multiple views, we build two experimental set-ups: Robustness to Occlusions and Recognition on Synthetic Data.

Robustness to Occlusions: Following the experimental protocol in [191], we train models on the training set of the NTU RGB+D 60 dataset and NTU RGB+120 and RGBD-Action-Completion-2016 dataset and tested under spatial and temporal occlusion conditions. For spatial occlusion, we evaluate the models using skeleton data with specific joints occluded. The occlusions correspond to the NO, LA, RA, WT, TH, and

TL, respectively. For temporal occlusion, we test models using skeleton sequences in which random blocks of frames are occluded, varying across nine levels i.e., 0 to 80 frames. By training on complete data and testing on incomplete data, we evaluate the robustness of the models. Each human body skeleton is presented via 25 key-joints, each described by 3D coordinates.

Recognition of Synthetic Data: To address the challenges posed by occlusions, we construct a synthetic occlusion dataset with diverse scenarios that simulate realistic conditions, categorizing occlusions into six types (NO, LA, RA, WT, TH, TL). Overall, 60% of datasets are randomly selected for occlusion experiments. The remaining 40% of the dataset consists of complete, unoccluded samples, as not all cases involve occlusion. The models are trained on synthetic datasets and tested under spatial and temporal occlusion conditions, consistent with the robustness experiments. This setup evaluates the model's ability to extract discriminative features from occluded skeleton data and handle activity recognition tasks. Uniform preprocessing is applied to ensure fair comparisons across all models.

5.3.3 Experimental Results

5.3.3.1 Unoccluded HAR Dataset

We analyze and contrast the performance assessment of our method with other techniques on unoccluded HAR dataset based on GCN in terms of their accuracy. We first train the model using six cases with complete skeleton sequence and test with incomplete skeleton sequence data for spatial occlusion as represented in Table 5.1, Table 5.3 and Table 5.5. These tables show the performance analysis of our model compared with other models on unoccluded datasets. With the observation, our model attains outstanding results when tested on occluded data with occluded body portions. ST-GCN [190] is the baseline approach of our model. Compared to the ASGCN [196], RA-GCN [189], STI-GCN [113], TCA-GCN [193], PDGCN [196], and MSFGCN [191], the proposed 2s PAST-GCN shows significant improvement, and also, the model's performance enhances overall. The 3s MSPAST-GCN demonstrates robustness against partial occlusion and its robustness increases with the addition of more model streams. For simulating the temporal occlusion, we randomly occluded a subsequence among the first 80 frames of the test data, adjusting the occlusion window to sizes of 10, 20, 30, 40, 50, 60,70, and 80 frames, respectively. Our 3s-PASTGCN model achieves better accuracy than other methods on occluded frames of varying lengths and highlights

the significant compensations in various situations with extensive temporal occlusion, as represented in Table 5.2, Table 5.4 and Table 5.6. Most models, including STI-GCN STI-GCN [113], and TCA-GCN [193], are not explicitly considered to handle occlusions, which is evident in their limited performance on occlusion benchmarks.

Table 5.1: Performance Assessment of Accuracy (%) for Spatial Occlusion on NTU-RGB+D 60 Dataset

Methods for Spatial Occlusion	No	LA	RA	WT	TH	TL	Avg.
ST-GCN [190]	80.7	71.4	60.5	50.2	62.6	77.4	67.13
2s-AGCN [196]	88.5	72.4	55.8	71.9	82.1	74.1	74.13
2s RA-GCN [189]	86.7	75.9	62.1	72.8	69.2	83.3	75.00
3s RA-GCN [189]	87.3	74.5	59.4	72.3	74.2	83.2	75.15
STI-GCN [113]	83.8	63.7	61.5	50.9	50.3	45.5	59.28
TCA-GCN [193]	88.4	64.8	60.2	72.4	59.4	70.8	69.33
2s PDGCN [196]	87.4	76.4	62.0	70.4	74.4	84.8	75.90
3s PDGCN [196]	87.5	76.0	62.0	73.0	75.4	85.0	76.48
MSFGCN [191]	90.2	88.3	74.7	78.6	75.6	87.4	82.46
2-Stream MSPAST-GCN	93.4	89.4	82.3	80.8	81.3	86.4	85.60
3-Stream MSPAST-GCN	94.6	87.3	78.9	76.5	82.4	90.2	84.98

Table 5.2: Performance Assessment in Accuracy (%) for Temporal Occlusion on Various Techniques on NTU-RGB+D 60 Dataset

Methods for Temporal Occlusion	0	10	20	30	40	50	60	70	80	Avg.
ST-GCN [190]	80.1	72.3	60.8	58.2	52.6	49.4	53.2	36.7	32.3	55.6
2s-AGCN [196]	84.3	76.6	72.4	68.7	60.6	48.5	40.8	38.4	26.4	57.41
2s RA-GCN [189]	86.5	83.1	66.4	62.6	58.3	40.5	39.6	26.3	20.2	53.72
3s RA-GCN [189]	84.8	83.9	76.4	66.3	53.2	38.5	42.2	39.0	23.4	56.41
STI-GCN [113]	88.8	70.4	51.0	38.7	33.8	28.0	23.4	17.4	14.2	40.63
2s PDGCN [196]	87.4	83.8	76.7	55.8	58.1	40.6	22.8	30.8	20.6	52.91
3s PDGCN [196]	87.5	83.9	76.6	66.7	53.9	38.0	36.2	42.1	24.3	56.57
MSFGCN [191]	90.2	89.2	84.2	81.2	76.3	74.3	72.7	68.3	65.1	77.94
2-Stream MSPAST-GCN	92.4	90.6	88.2	86.3	82.1	78.7	72.3	67.4	66.4	80.48
3-Stream MSPAST-GCN	96.6	94.6	91.2	90.2	87.8	89.3	84.3	78.4	75.1	87.50

5.3.3.2 Occluded Synthesized Dataset

We analyze & contrast the performance of our method with other approaches on synthesized occluded datasets produced from CS and CSET views of the NTU-RGB+D 60 and 120 datasets, respectively, and the RGBD-Action-Completion-2016 dataset. Our

Table 5.3: Performance Assessment in Accuracy (%) for Spatial Occlusion on Various Techniques on NTU-RGB+D 120 Dataset

Methods for Spatial Occlusion	No	LA	RA	WT	TH	TL	Avg.
ST-GCN [190]	80.3	71.4	60.5	50.2	62.6	77.4	67.06
2s-AGCN [196]	85.5	72.4	55.8	71.9	82.1	74.2	73.70
2s RA-GCN [189]	86.7	75.9	62.1	70.8	69.2	80.3	74.16
3s RA-GCN [189]	87.3	74.5	59.4	68.3	74.2	79.2	73.81
STI-GCN [113]	63.7	42.6	41.5	30.9	38.3	45.1	43.68
2s PDGCN [196]	87.5	76.0	62.0	70.1	75.4	82.0	75.50
3s PDGCN [196]	90.1	88.3	74.7	78.6	75.6	87.4	82.45
MSFGCN [191]	93.4	89.4	82.3	80.8	81.3	86.4	85.60
2-Stream MSPAST-GCN	94.6	87.3	78.9	76.5	82.4	89.2	84.81
3-Stream MSPAST-GCN	92.3	89.4	73.8	77.9	82.1	83.4	83.15

Table 5.4: Performance Assessment in Accuracy (%) for Temporal Occlusion on Various Techniques on NTU-RGB+D 120 Dataset

Methods for Temporal Occlusion	0	10	20	30	40	50	60	70	80	Avg.
ST-GCN [190]	83.8	76.9	67.9	66.6	64.5	60.5	58.9	56.4	48.3	64.86
2s-AGCN [196]	84.5	75.9	66.9	64.1	60.2	58.4	54.4	49.8	38.6	61.42
2s RA-GCN [189]	84.0	77.1	70.5	71.7	66.6	62.2	56.4	53.4	46.1	65.33
3s RA-GCN [189]	85.8	82.4	80.1	79.8	73.4	62.7	58.2	49.4	35.2	67.44
2s PDGCN [196]	85.9	78.8	74.6	68.2	54.8	50.8	49.4	40.3	36.4	59.91
3s PDGCN [196]	84.8	82.7	79.6	70.5	66.7	54.6	42.6	34.8	29.7	60.67
MSFGCN [191]	85.1	79.6	74.9	69.7	60.3	59.6	56.5	42.4	36.1	62.68
2-Stream MSPAST-GCN	89.4	84.2	80.8	79.3	74.6	70.8	69.4	65.6	60.4	74.94
3-Stream MSPAST-GCN	90.5	89.0	86.0	84.0	80.4	79.8	77.4	76.0	68.0	81.23

comparison methods' experimental test accuracy for spatial occlusion and temporal occlusion with other approaches on skeleton-based GCN are concise in Table 5.7 and Table 5.8. We incorporated a preprocessing module across all models to improve the occlusion and trained them on a synthesized occlusion dataset we developed. Compared to the outcomes in Table 5.1, Table 5.3 and Table 5.5, the overall model's accuracy improved by an average of 6%, as shown in Table 5.7, validating the effectiveness of our training strategy. While there is a slight decrease in recognition accuracy compared to the average output in Table 5.2 and Table 5.3, but again increases in Table 5.4, this is comprehensible given that synthesized dataset lacks samples for temporal dimensions and mainly emphasizes spatial dimensions occlusion which may have affected the learning of temporal features. However, this minor reduction in temporal accuracy is acceptable, considering the significant gains in spatial occlusion accuracy.

Table 5.5: Performance Assessment in Accuracy (%) for Spatial Occlusion on Various Techniques on RGBD-Action-Completion-2016 Dataset

Methods for Spatial Occlusion	No	LA	RA	WT	TH	TL	Avg.
ST-GCN [190]	78.4	71.4	60.5	50.2	62.6	77.4	66.75
2s-AGCN [196]	88.5	72.4	55.8	71.9	82.1	74.1	74.13
2s RA-GCN [189]	86.7	75.9	62.1	72.8	69.2	78.3	74.16
3s RA-GCN [189]	87.3	74.5	59.4	72.3	74.2	80.2	74.65
STI-GCN [113]	88.8	23.7	21.5	20.9	20.3	45.5	36.78
2s PDGCN [196]	87.5	76.0	62.0	73.0	75.4	85.0	76.48
3s PDGCN [196]	90.2	88.3	74.7	78.6	75.6	87.4	82.46
MSFGCN [191]	93.4	89.4	79.3	80.8	81.3	86.3	85.08
2-Stream MSPAST-GCN	94.6	89.3	91.9	90.5	93.4	92.5	92.03
3-Stream MSPAST-GCN	95.6	91.2	90.8	89.8	88.6	89.2	90.86

Table 5.6: Performance Assessment in Accuracy (%) for Temporal Occlusion on Various Techniques on RGBD-Action-Completion-2016 Dataset

Methods for Temporal Occlusion	0	10	20	30	40	50	60	70	80	Avg.
ST-GCN [190]	78.2	71.1	60.4	51.2	52.6	50.4	49.8	46.3	36.6	55.17
2s-AGCN [196]	84.8	82.7	79.6	70.5	66.7	54.6	42.6	34.8	29.7	60.66
2s RA-GCN [189]	85.6	80.6	78.9	69.9	60.5	59.6	57.6	44.8	30.2	63.07
3s RA-GCN [189]	84.8	81.3	80.6	78.4	72.4	60.2	56.6	48.4	34.4	66.34
STI-GCN [113]	85.8	84.4	82.1	79.8	73.4	62.7	58.2	49.4	35.2	67.88
2s PDGCN [196]	84.8	82.7	79.6	70.5	66.7	54.6	42.6	34.8	29.7	60.66
3s PDGCN [196]	85.1	79.6	74.9	69.7	60.3	59.6	56.5	42.4	36.1	62.68
MSFGCN [191]	84.8	81.3	80.6	78.4	72.4	60.2	56.6	48.4	34.3	66.33
2-Stream MSPAST-GCN	90.5	76.0	62.0	73.0	75.4	85.0	90.5	76.0	62.0	76.71
3-Stream MSPAST-GCN	92.2	88.3	74.7	78.6	75.6	87.4	92.2	88.3	74.7	83.55

5.3.4 Ablation Studies

To evaluate the contributions of each module in the MSPAST-GCN model, we conducted an ablation study by systematically removing three key components: IIM, PSI, and PAST-GCN, which are designed to improve occlusion robustness. The results reveal distinct performance degradation patterns when any of these components are removed, emphasizing their critical role in feature learning and occlusion handling. The removal of IIM resulted in the most significant accuracy drop, approximately 13.6% for spatial & 27.2% for temporal occlusions, as the model lost its ability to handle missing key joints effectively. The primary function of IIM is to expose the model to occlusion scenarios during training, enabling it to infer missing key joint information. Without

Table 5.7: Performance Assessment in Accuracy (%) for Spatial Occlusion on Various Techniques on Synthesized Occlusion Dataset

Methods for Spatial Occlusion	No	LA	RA	WT	TH	TL	Avg.
ST-GCN [190]	76.9	70.2	61.5	52.2	62.6	77.4	66.80
2s-AGCN [196]	84.5	75.9	66.9	79.1	81.0	72.2	76.51
2s RA-GCN [189]	88.5	72.4	55.8	71.9	82.1	74.1	74.13
3s RA-GCN [189]	86.7	73.9	62.1	70.2	69.2	82.1	74.03
STI-GCN [113]	87.3	74.5	59.4	72.3	74.2	83.2	75.15
2s PDGCN [196]	89.5	80.2	74.0	72.8	75.4	82.4	79.05
3s PDGCN [196]	90.2	88.3	74.7	78.6	75.6	74.4	80.30
MSFGCN [191]	87.4	76.4	62.0	70.4	74.4	84.8	75.90
2-Stream MSPAST-GCN	90.2	86.0	82.0	76.0	74.4	77.0	80.93
3-Stream MSPAST-GCN	92.4	84.3	80.7	79.6	76.6	79.4	82.16

Table 5.8: Performance Assessment in Accuracy (%) for Temporal Occlusion on Various Techniques on Synthesized Occlusion Dataset

Methods for Temporal Occlusion	0	10	20	30	40	50	60	70	80	Avg.
ST-GCN [190]	84.8	82.7	79.6	70.5	66.7	54.6	42.6	34.8	29.7	60.61
2s-AGCN [196]	85.6	80.6	78.9	69.9	60.5	59.6	57.6	44.8	30.2	63.07
2s RA-GCN [189]	84.8	81.3	80.6	78.4	72.4	60.2	56.6	48.4	34.4	66.34
3s RA-GCN [189]	85.8	84.4	82.1	79.8	73.4	62.7	58.2	49.4	35.2	67.88
2s PDGCN [196]	85.9	78.8	74.6	68.2	54.8	50.8	49.4	40.3	36.4	59.91
3s PDGCN [196]	88.6	84.1	80.2	74.4	70.8	61.6	42.6	40.6	38.2	64.56
MSFGCN [191]	92.3	88.1	81.7	72.3	70.4	66.8	54.2	48.4	42.6	68.53
2-Stream MSPAST-GCN	94.8	93.2	90.1	88.7	89.3	84.2	80.6	76.7	72.3	85.54
3-Stream MSPAST-GCN	96.4	93.8	89.2	87.9	85.4	83.1	79.4	74.3	71.8	84.54

IIM, the model becomes over-reliant on fully visible skeletons, significantly reducing its generalizability to occluded environments. Similarly, removing PSI caused a 2.4% drop in spatial & and 8.2% drop in temporal occlusion accuracy. PSI prevents the model from assigning excessive importance to visible key joints, promoting balanced feature learning. Without PSI, the model overfits to available key joints, diminishing its adaptability to missing or occluded parts. The removal of PAST-GCN led to an 11.5% accuracy drop for spatial & 19.8% drop for temporal occlusions. PAST-GCN captures critical spatial-temporal dependencies between key joints, allowing the model to preserve motion continuity. Without it, the model faces issues in maintaining temporal coherence, particularly during long-term occlusions. Furthermore, the combined effect of removing both IIM and PSI resulted in compounded performance degradation, indicating their complementary roles in occlusion simulation and adaptive score

inhibition. Fig. 5.6 provides a visual representation of the impact of each module on overall accuracy. It highlights that IIM contributes the most to temporal occlusion robustness, resulting in a -7.4% accuracy drop when removed. Similarly, PAST-GCN plays a critical role in preserving motion continuity, reducing accuracy by -8.6% for spatial occlusions and -11.1% for temporal occlusions. Lastly, PSI enhances adaptive feature learning, causing smaller performance reductions of -2.4% for spatial occlusions and -8.2% for temporal occlusions when removed. These results confirm that each module has a distinct and essential role in the proposed architecture. Their integration significantly enhances the model's ability to handle both spatial and temporal occlusions, ultimately improving overall classification accuracy.

The evaluation is performed under both spatial occlusion (e.g., partial blockage of a subject) and temporal occlusion (e.g., sudden interruptions in motion sequences). We train a single-stream and multi-stream GCN fusion network to determine the efficacy of the proposed model, and the comparison resultant accuracy for each occluded portion among them is described in Table 5.9. We must utilize multi-scale motion representation to evaluate the weight bias while observing that the multi-stream fusion net performs markedly improved results than the single-stream network. We calculate the actual performance of MSPAST-GCN module in spatial-temporal occlusion scenarios while retaining their principal originality. Fig. 5.7 depicts the training validation for each baseline method with our proposed framework. The graphical representation clearly shows the training accuracies of all baseline methods and offers the superior performance of MSPAST-GCN. Additionally, Fig. 5.7a shows a comparison on the NTU RGB+60 dataset, where the MSPAST-GCN method achieves the best accuracy at the 200th epoch. Similarly, in Fig. 5.7b for the NTU RGB+120 dataset, the proposed method achieves a better accuracy score over the 200th epoch. Fig. 5.7c for the RGBD-Action-Completion-2016 dataset, our method performs best accuracy at 200th epochs. Lastly, in Fig. 5.7d for the Synthesized Occlusion dataset, the MSPAST-GCN method attains the finest training accuracy all over the training phase of other methods.

Table 5.9: Ablation Studies on CS Benchmark of NTU-RGB+D 60

Technique	NO	LA	RA	WT	TL	TH	Avg. Gain
Single-Stream Network	84.2	78.2	64.9	68.2	80.2	63.5	73.20
Multi-Stream Network	86.8	86.4	72.8	80.2	86.4	79.8	82.06
MSPAST-GCN (Proposed)	88.1	84.8	76.6	81.3	86.3	80.8	82.98

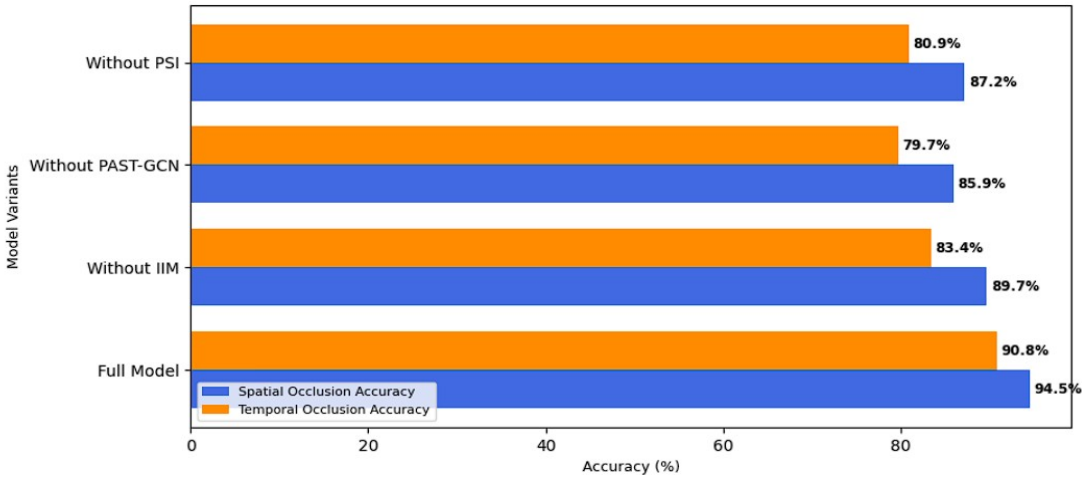
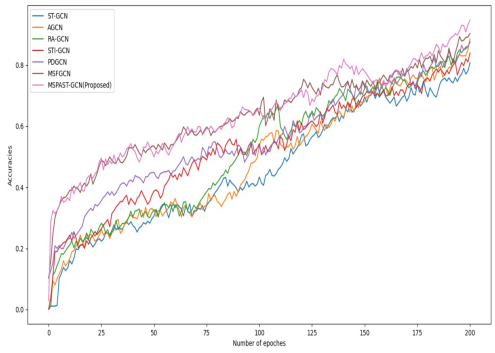
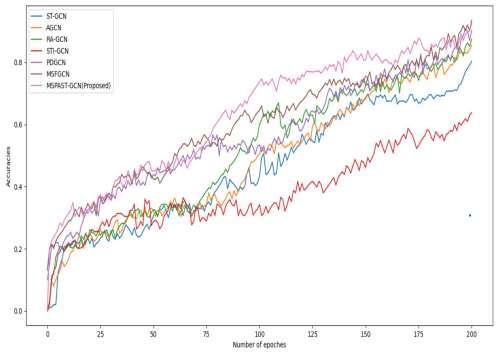


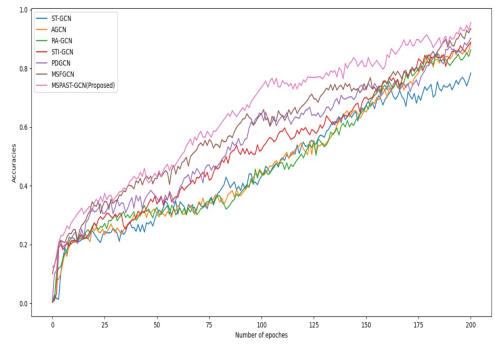
Fig. 5.6: Impact of Different Modules on Accuracy



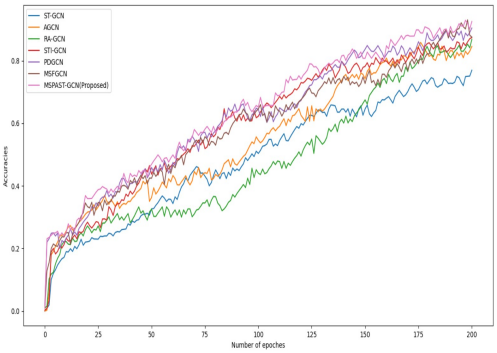
(a) Training Accuracies for NTU RGB+D 60 Dataset



(b) Training Accuracies for NTU RGB+D 120 Dataset



(c) Training Accuracies for RGBD-Action-Completion-2016 Dataset



(d) Training Accuracies for Synthesized Occlusion Dataset

Fig. 5.7: Trade-off curves of the proposed MSPAST-GCN with other verified baseline methods over various HAR and synthesized datasets.

5.3.4.1 Quantitative Analysis

Table 5.10 presents the performance metrics when each module is removed individually. The full model achieves the highest accuracy, whereas removing any component

leads to a significant decline in performance. From the result, it is evident that the IIM module contributes significantly to precision, while the PSI and PAST-GCN modules play crucial roles in recall and spatial awareness under occluded environments.

Table 5.10: Ablation Studies on Occluded Synthesized Dataset

Model Variation	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
MSPAST-GCN (full-model)	94.5	92.3	90.8	91.5
Without IIM	89.7	86.1	83.4	84.7
Without PSI	87.2	84.5	80.9	82.6
Without PAST-GCN	85.9	83.2	79.7	81.4

5.3.4.2 Qualitative Analysis: Visualizations of Key-Joint Activation

Fig. 5.8 shows the detailed visualization of key-joint activation patterns in each stream of ST-GCN, equivalent to activities such as removing spectacles, fighting, kicking, and suspicious walking. The activation status of each key joint is carefully mapped according to their activity prediction, allowing us to analyze inconsistencies and validate the robustness of the model while detecting human activity. The activated key joints of the body are represented by red circles, while green circles indicate the remaining occluded key joints. Compared to the SOTA methods, we observe that the proposed model focuses more effectively on the crucial key joints closely associated with the specific activities. For instance, in Fig. 5.8a, the activity shows for removing spectacles, which is deceptive that hand motion activities are more critical than foot motion activities. The model focuses on hand key joints per the six occluded cases and then works on streams. We also perceive significant variations in the activated key-joints across the five frames of the streams. This can be clarified by the IIMs module, which inhibits the input data matching to previously activated positions in earlier streams, similar to the cases in Fig. 5.8b, 5.8c, and 5.8d. This navies the model to stimulate other potentially valuable key-joints, thus improving the MSPAST-GCN model's robustness against occlusions.

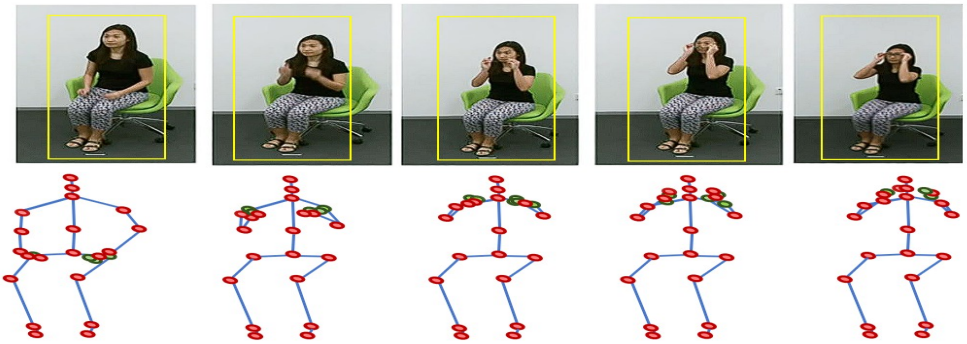
5.3.5 Discussion

Despite the significant improvements achieved by MSPAST-GCN in occluded skeleton-based human activity recognition, certain limitations remain. The model misclassify in cases of extreme occlusions, such as full-body occlusions or overlapping individuals, where key joints remain invisible for extended periods. Additionally, similar activities involving subtle motion differences such as drinking from a cup vs eating or picking up

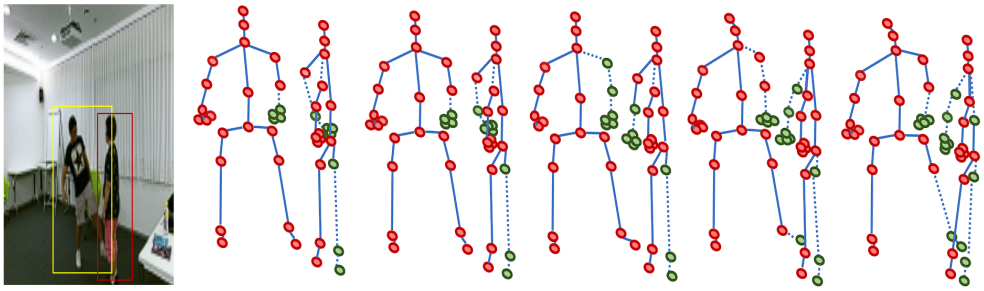
an object vs throwing, can lead to misclassifications due to occlusion-induced loss of discriminative features. The multi-stream architecture, while enhancing feature extraction, increases computational complexity, making real-time deployment on resource-constrained devices challenging. Although the model generalizes well across benchmark datasets, real-world occlusions may differ, necessitating further fine-tuning. In future work, we will focus on improving real-time efficiency, increasing adaptability across diverse occlusion scenarios, and integrating real-world occlusion cases for better generalization.

5.4 Summary

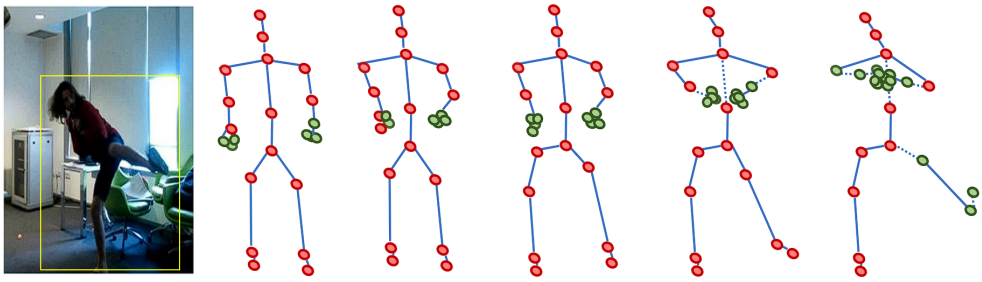
This chapter presented a multi-stream part-aware occluded skeleton-based GCN for HAR, utilizing inhibition training in different streams to handle occlusions across various conditions, thus improving accuracy in such occlusion scenarios. This method enables the learning of local and global features, enhancing performance and offering a practical solution, similar to ST-GCN. The proposed module consists of an input inhibition skeleton sequence, part-aware ST-GCN for graph construction, and a predicted score inhibition module. When evaluated under different occlusions, the proposed model consistently outperforms SOTA methods across multiple experimental configurations. The ablation study further validates the effectiveness of each module in the model. Additionally, visualizations demonstrate the model's ability to accurately recognize deeply occluded samples, highlighting its strength in handling occlusions across both spatial and temporal dimensions. This research introduces an optimized training approach that significantly improves average recognition efficiency on HAR datasets of about 6% of achievement, surpassing previous work's performance.



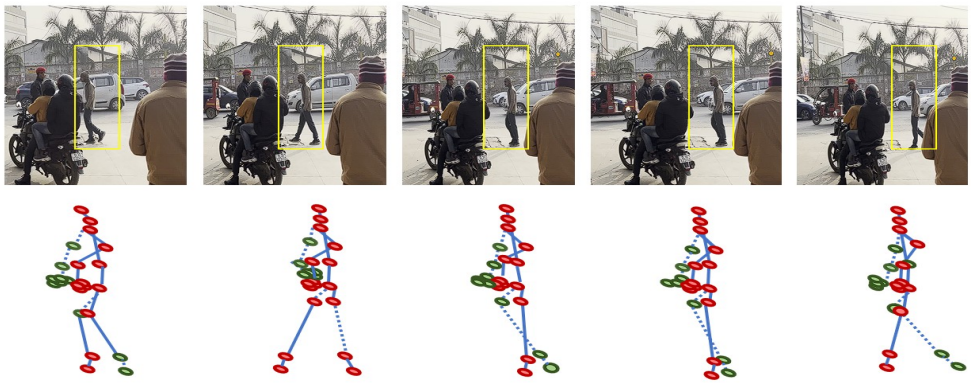
(a) Removing Spectacle



(b) Fighting (Multi-Person)



(c) Kicking



(d) Suspicious Walk

Fig. 5.8: Visualization of Body Key-joint Activation for the proposed model with their prediction results on (a) NTU-RGB+D 60, (b) NTU-RGB+D 120, (c) RGBD-Action-Completion-2016, and (d) occluded synthesized dataset. Note: Red circles indicate activated key joints, while green circles represent occluded key joints.

Chapter 6

Activity Recognition in Dynamic Environments Using Image Enhancement and Vision Transformers with DETR

This chapter introduces a novel low-light enhancement technique tailored to improve the performance of HAR systems under challenging illumination conditions. The proposed method enhances illumination intensity, contrast, and color consistency while preserving spatial details in low-light images. The model's effectiveness is validated in both normal and low-light environments through extensive quantitative and qualitative analysis, demonstrating superior performance compared to existing approaches.

6.1 Introduction

In dynamic environments, human activity refers to movements and behaviors performed under changing conditions like varying backgrounds, lighting, or motion. Images taken from video datasets in poor light or dark environmental conditions somehow destructively affect human observation. It also degrades the effectiveness of computer vision tasks of some applications, such as sports prediction, elderly fall detection, patient monitoring, smart surveillance systems, facial recognition, etc. [130][197]. However, low brightness and contrast in monitoring images hinder target recognition in dark or low-light conditions. These issues lead to blurred targets, obscuring feature extraction and analysis in advanced visualization. Additionally, regions near light sources are prone to dust-induced overexposure and whitening, further degrading image quality. The coexistence of overexposure and underexposure in images complicates the task of image enhancement.

DL-based LLIE has significant social benefits, improving safety and efficiency in critical environments like night-round in the military, safety surveillance in park-

ing at night, underground coal mines. These methods surpass traditional approaches in accuracy but face challenges such as the impracticality of collecting light-paired datasets, limitations of synthetic data, and the trade-off between enhancement quality and real-time detection. By overcoming these challenges, such progress in LLIE can have profound social benefits, including enhanced safety in hazardous conditions, improved surveillance for public security, and better access to critical resources in low-visibility environments. The LLIE is crucial for enlightening video monitoring quality and recognition capabilities. It involves transforming low-light images into natural standard illumination images by addressing three key steps: Diminishing noise and artifacts, keeping edges and surface details intact, and recovering authentic brightness, color, and image pixels. However, the process frequently leads to loss of information like brightness, color, structure, and contrast, while also amplifying noise and haze in the enhanced images. We developed a LLIE method collective with Vision Transformer (ViT) with DETR for monitoring in the public sector. The key contributions are:

- A low-LLE technique is proposed to enhance illumination intensity, contrast, and color dependability of input images under low-illumination environments while preserving the spatial specifics of the images.
- To verify the strength of the proposed model, a comprehensive study is executed across three challenging datasets, i.e., SCIE, LOLO-V1, and ARID.
- Furthermore, the prediction performance of the proposed model is evaluated under both normal and low-lighting environments.
- The experimental assessment is also measured from both quantitative and qualitative perspectives, showing that the framework outperforms existing networks in HAR.

6.2 Proposed Methodology

We introduced a zero-reference low-light enhancement approach, based on the zero-reference learning method [198] to retain the specifics of the input image. Unlike other deep learning techniques, our method adjusts only the pixel signals instead of regenerating them. The outline for proposed model for LLIE is shown in Fig. 6.1. and comprises with two phases: (i) local pixel-level image enhancement (ii) transformer-based global adjustment.

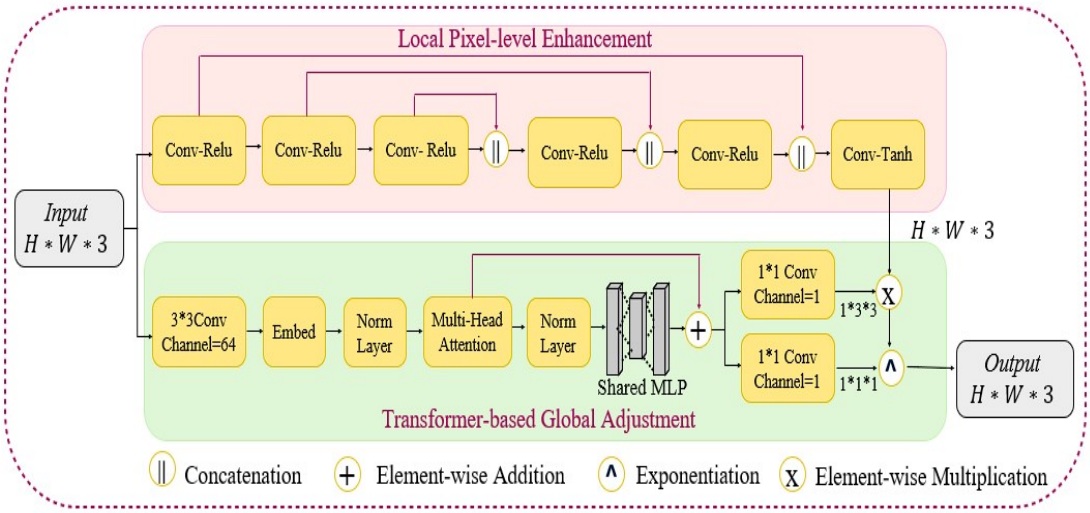


Fig. 6.1: Pipeline of Proposed Model for LLIE

6.2.1 Phase 1: Local pixel-level image enhancement

We emphasize assessing pixel-level image illumination mapping to correct illumination effects within image. We maintain the input resolution to preserve detailed information and prevent negative impacts on subsequent detection results rather than using down-sampling trailed by up-sampling. The local pixel-level image enhancement module comprises 7 convolutional layers & final layer utilizes Tanh activation function, while the remaining layers adopt the ReLU activation function, drawing inspiration from [199]. In contrast to their earlier work, we utilize a reciprocal-based mapping function as a substitute of a quadratic iterative function. The quadratic curve-based illumination mapping function can be represented as:

$$L_E(I_X;\alpha) = I_X + \alpha.I_X(1 - I_X), \quad \alpha \in [-1, 1] \tag{6.1}$$

where $L_E(I_X;\alpha)$ represents the improved type of input I_X of the image with x as pixel coordinates . $\alpha \in [-1, 1]$ indicates the trainable curve parameter obtained by deep network training. Every specific pixel is normalized to interval $[0, 1]$. The light mapping curvature depends on the quadratic iteration function for higher-order curves to enable versatile adjustment to image pixels can be defined as:

$$L_{E_n}(I_X) = L_{E_{n-1}}(I_X) + A_n L_{n-1}(I_X)(1 - L_{n-1}(I_X)) \tag{6.2}$$

where n indicates the iterations that control the curving, and A represents a parameter map that matches the sizes of I_X.

6.2.2 Phase 2: Transformer Global Adjustment in Image

We developed a global adjustment module utilizing a transformer to address the limitations of local enhancement by capturing global interactions among specific pixels with their environments. The Transformer Global Adjustment (TGA) in image branch primarily comprises a multihead attention module and a Multi-Layer Perceptron (MLP) module. Its process employs layer normalization to normalize implicit layers and accelerate convergence. The input image is first processed with a 3×3 convolution, producing high-dimensional, low-resolution features. This design reduces computational costs while enabling global feature extraction. For feature encoding, the 3×3 convolution produces a feature map F_m of size $64 \times h \times w$. Here, every pixel of F_m is treated as a token, flattened into a sequence X_T of size $(h \times w) \times 64$. A linear transformation encodes X_T as input to the multihead attention module. The encoded sequence X_T is projected using learnable weights W^p, W^q, and W^r to form P, Q, and R. The calculation process can be signified as:

$$P = X_T W^p, \quad Q = X_T W^q, \quad R = X_T W^r \tag{6.3}$$

$$A_{tt}(\text{attention}) = \text{softmax}\left(\frac{\sqrt{PQ^T}}{d_k}\right) R \tag{6.4}$$

Multi-head attention processes P, Q, and R in multiple subspaces, concatenates the results, and applies a linear transformation:

$$\text{MultiHead}(P, Q, R) = \text{Concat}\left(\text{head}_1, \text{head}_2, \ldots, \text{head}_n\right) \tag{6.5}$$

For color and gamma adjustments, the multihead attention results in a F_m of specific size processed by an MLP to compute a color transformation matrix $W_{(ci,cj)}$ for RGB adjustment. This matrix is 3×3, and the gamma value γ is utilized for nonlinear global illumination correction. The final global adjustment is expressed as:

$$G_{\text{adj}}(.) = \max\left(\sum_{cj} W_{(ci,cj)}(.), \epsilon\right)^{\gamma}, \quad ci, cj \in \{r, g, b\} \tag{6.6}$$

Here, ci, cj indicates the colour transformation matrix of size 3×3, γ represents the exponential value of gamma illumination correction, ϵ implies a minimum non-

negative value, which takes the constant $\epsilon = 1e^{-8}$ in the entire setup.

The complete process of the low-LLE method can be stated as:

$$G_{\text{adj}}(L(I_X)) = \left(\max \left(\sum_{cj} W_{(ci,cj)}(L(I_X)), \epsilon \right) \right)^{\gamma}, \quad ci, cj \in \{r, g, b\} \tag{6.7}$$

6.2.3 Loss Function for Low-Light Image Enhancement

The LLIE losses are categorized into three losses, i.e., exposure control loss, spatial consistent loss, and color consistent loss function. These are formulated as:

$$L_{\text{exp}} = \frac{1}{m_1} \sum_{i=1}^{m_1} |Y_m - \epsilon| \,; \quad L_{\text{spa}} \quad = \frac{1}{m_2} \sum_{i=1}^{m_1} \sum_{j \in \Omega(i)} \left(|Y_i - Y_j| - |I_i - I_j| \right)^2 \tag{6.8}$$

$$L_{\text{col}} = \sum_{\forall (p,q) \in \epsilon}^{m_1} \left((J^p - J^q)^2 \right), \quad \epsilon = \{(R,G),(R,B),(G,B)\} \tag{6.9}$$

here m_1 and m_2 represent non-overlapping localized areas, Y_m is their intensity, and $\epsilon = 0.6$. $\Omega(i)$ denotes neighboring regions of i, while Y & I are the average intensities of improved and input images. The color constancy loss ensures consistent enhanced colors, with J^p and J^q representing average intensities of channels p and q. Fig. 6.2 shows the process for computing the feature similarity loss. F_m of the input and enhanced images are extracted using the ViT backbone with DETR. These F_m undergo GAP, and the feature similarity loss is computed as follows:

$$L_{\text{fs}} = \text{GAP}(F_{\text{IN}}) \cdot \log \left(\frac{\text{GAP}(F_{\text{EN}})}{\text{GAP}(F_{\text{IN}})} \right) \tag{6.10}$$

where GAP(F_{IN}) indicates the feature global pooling of the (F_{IN}), and GAP(F_{EN}) indicates the feature global pooling of the (F_{EN}). Thus, the final LLIE loss is represented as:

$$L_{\text{total}} = L_{\text{fs}} + L_{\text{col}} + L_{\text{spa}} + L_{\text{exp}} \tag{6.11}$$

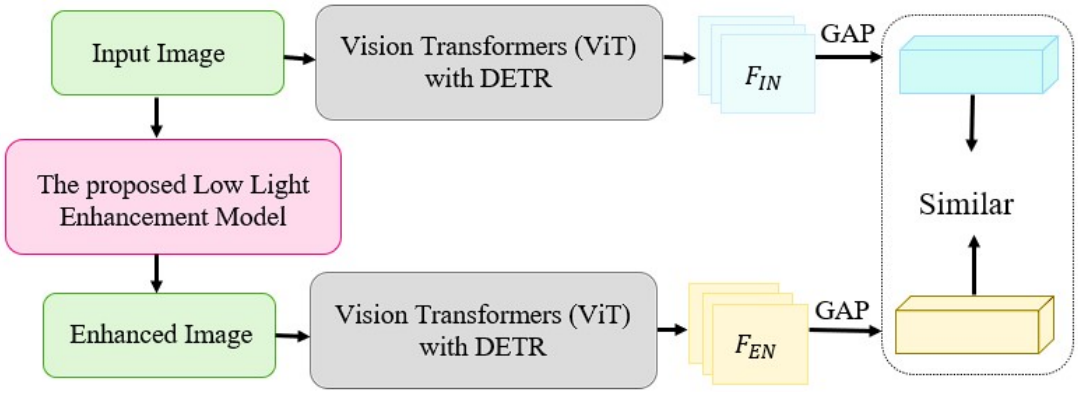


Fig. 6.2: Block diagram of feature similarity loss.

6.3 Experimentation and Results

6.3.1 Setup Details and Datasets

We implement our proposed framework on PyTorch 1.12.2, Python 3.9, CUDA 11.6, NVIDIA RTX 3040Ti GPU. To optimize the model we utilize the ADAM optimizer with default parameters, setting the learning rate 1×10^{-4}, batch size 8, momentum to 0.9, and training for 200 epochs. The decay coefficient is set to 5×10^{-5}. We resized the training images to resolution 224×224.

SICE Dataset (Single-Image Contrast Enhancement) [197]: This dataset is designed for evaluating image enhancement algorithms. It includes high-resolution images with diverse lighting conditions, ranging from underexposed to overexposed scenarios. It consists of 589 multiple exposure pictures of indoor and outdoor surroundings. For training the LLIE module, we selected 3022 images with diverse exposure levels.

LOL-V1 Dataset (Low-Light Dataset Version-1 [200]): This dataset features paired low-light and well-exposed reference images. It includes diverse indoor and outdoor scenes, making it ideal for evaluating and training algorithms to enhance visibility and detail in low-light surroundings. The dataset consists of 500 pairs low-light and normal-light images, each with 400×600 resolution. A total of 485 pairs are designated for training, while the remaining 15 pairs for testing.

ARID Dataset (Action Recognition in the Dark) [201]: This dataset features real-world videos with detailed annotations, making it valuable for developing and testing

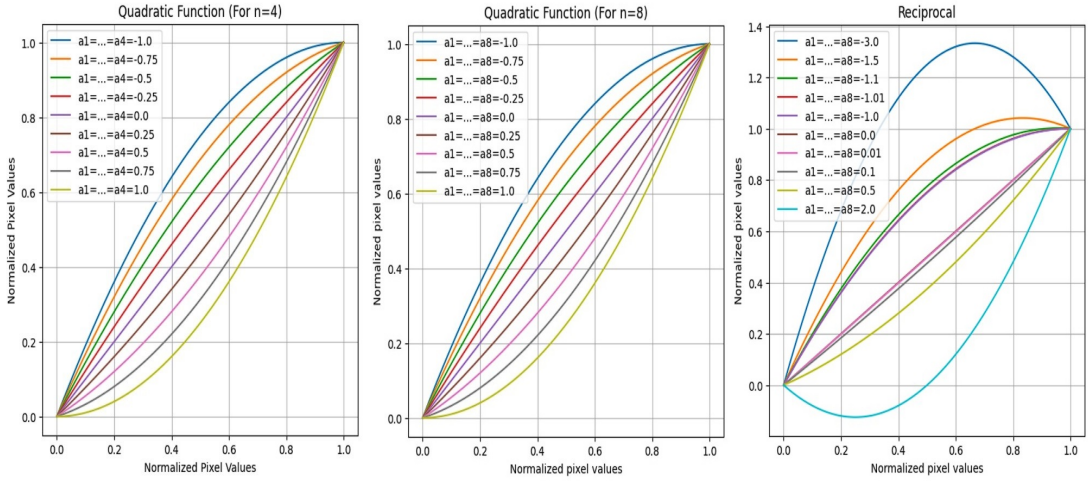


Fig. 6.3: The function for adaptive light mapping curvature having varying parameters. Curves a and b correspond to the quadratic iterative function, for 'n' to 4 & 8 times, respectively. Curve c represents the light mapping on the reciprocal function.

models for surveillance and security applications in dark environments. The dataset includes 11 action classes with videos shot at night in both indoor and outdoor settings, at 30 fps and 320×240 resolution.

Table 6.1: Quantitative Results of SOTA on SCIE, LOL-V1, ARID Datasets.

Techniques	SICE				LOL-V1				ARID			
	PSNR	SSIM	NIQE	BRIS	PSNR	SSIM	NIQE	BRIS	PSNR	SSIM	NIQE	BRIS
Retinex-Net[200]	16.70	0.59	15.20	0.52	16.70	0.65	14.20	0.50	18.10	0.60	15.20	0.51
EnlightenGAN [202]	19.83	0.62	9.39	0.55	19.83	0.69	10.39	0.48	18.83	0.66	10.23	0.52
Zero-DCE[203]	15.77	0.53	10.22	0.48	15.77	0.78	12.02	0.52	19.87	0.64	10.11	0.49
DRBN-YOLOv5s [199]	18.23	0.76	10.29	0.52	18.23	0.77	10.29	0.55	20.89	0.77	9.46	0.54
Proposed	21.70	0.77	10.01	0.54	21.70	0.77	10.01	0.56	20.88	0.79	9.03	0.55

6.3.2 Quantitative Result Analysis

This section presents an exhaustive quantitative analysis of the proposed and SOTA techniques on the SCIE, LOL-V1, and ARID datasets. We used two categories of four metrics, i.e., full references and no-reference, and the metrics are PSNR, SSIM, NIQE, and BRISQUE, for image quality evaluation. PSNR and SSIM evaluate image enhancement with high values represents better quality and similarity to the ground truth. For ground-truth-free evaluation, BRISQUE (↑ is better) and NIQE (↓ is better) assess perceptual quality and naturalness. Table 6.1 summarizes the quantitative results of SOTA methods.

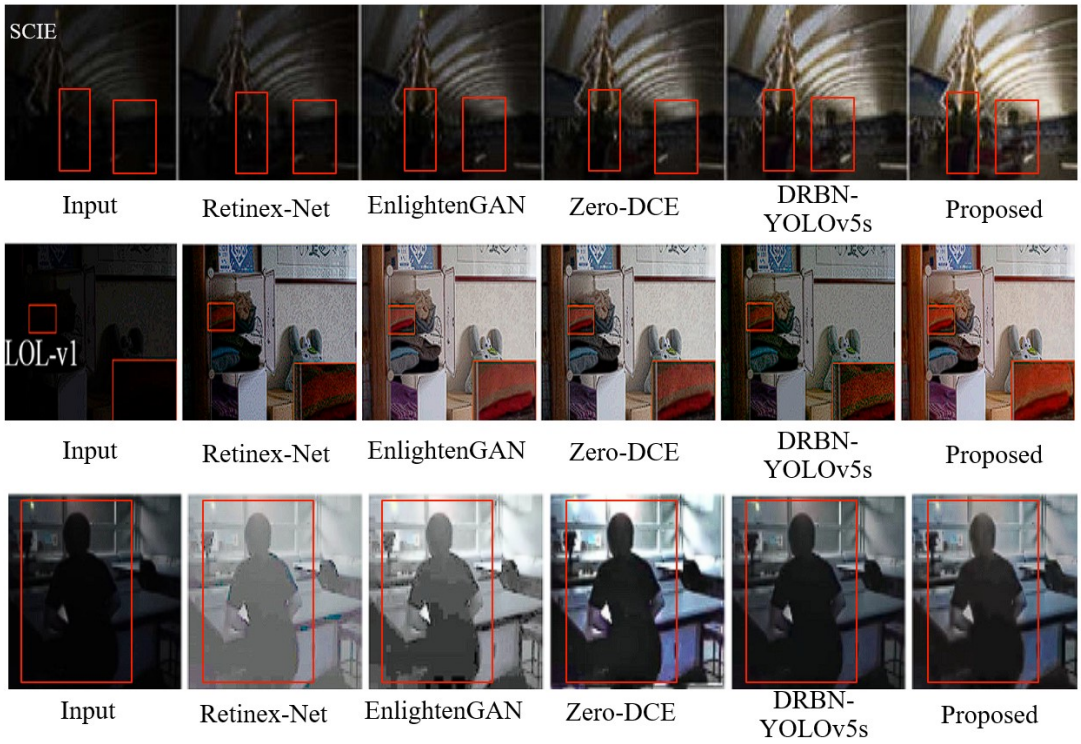


Fig. 6.4: Visual comparisons of LLE methods, starting from the input image to results of various techniques, with red-boxed areas zoomed in on the human as an object.

6.3.3 Qualitative Result Analysis

The image enhancement results across datasets are illustrated in Fig. 6.4 Retinex-Net enhanced brightness but caused texture distortion and blurring. Methods like Zero-DCE, EnlightenGAN, DRBN-Yolov5s, and our proposed technique performed better overall. Zero-DCE showed slight color shifts, and EnlightenGAN under-improved dark areas. Our proposed technique achieved superior illumination and contrast, preserving color and texture without noticeable bias or distortion.

6.4 Summary

This chapter introduces a zero-reference low-light image enhancement approach to improve visualization and recognition in dark or low-light environments. This approach follows end-to-end train beyond the need for reference images. It can be achieved by framing the LLIE chore as an image-specific curvature estimation problematic while developing a set of differentiable non-reference losses to optimize the image enhancement process. The experimental outcomes verify the proposed approach suggestively improves the intensity of the image, contrast, and recognition rate while preserving

their original color and avoiding texture distortion. Moreover, the method eliminates the need for manual parameter adjustment and performs effectively on both overexposed and under-illuminated images. It significantly strengthens the ability and efficiency of individuals' safety and security monitoring in low-light environments.

Chapter 7

MV-DBiLSTM: An Enhanced Human Activity Recognition for Smart Surveillance Systems Using a Deep BiLSTM Framework

In this chapter, we present a novel machine learning framework, MV-DBiLSTM, for human activity recognition. The proposed model integrates MobileNetV2 for efficient spatial feature extraction and a deep bi-LSTM to capture both short-term and long-term temporal dependencies in activity sequences. This dual-stream approach enables the model to recognize complex and dynamic human actions with improved accuracy.

7.1 Introduction

Smart surveillance focuses on recognizing normal vs. abnormal behavior, detecting suspicious activities (e.g., trespassing, aggression), tracking individuals over time for event analysis or alerts. For example, the initial action of fast walking and jogging in a playground may appear similar. That difference becomes clearer once we see them across a series of frames, where the interaction of the human body with its environment can be captured through their gestures and movements. The previously trained model on data might become obsolete with the availability of new data because, in non-stationary video data streams, the data changes over time. To solve this problem, one needs to adapt to the innovative data distribution for heterogeneous static surroundings. Lobo et al. [204] formulated a self-learning optimization problem utilizing bio-based optimization techniques to handle drift heterogeneity. Krawczyk et al. [205] proposed a technique called weighted single class SVM, that uses a modified version to improve the efficiency of static streaming data which adapts new data streams and understands them by relearning a portion of its parameters. Abdallah et al. [206] provided an extensive review of AR for online data streaming. Nevertheless, detecting human activities

from the huge volume of surveillance streaming data is a challenging problem owing to high-dimensional characteristics, changing viewpoint, motion, background clutter, occlusion, and varying illumination conditions [207, 208, 209]. Some automatic feature learning techniques based on neural networks become more adaptive due to their supervised learning nature. These approaches can directly compute the features from raw input information with the weights and biases of learned network. In this, firstly, the initial layers capture the low-level local features like edges and textures, and then the next layers extract more complex, global features that represent high-level semantic information. This hierarchical structure allows CNNs to recognize intricate patterns and provide an overall understanding of the image. Recently, several authors have studied CNN-based techniques, Transfer Learning, and Transformer approaches to maximize performance on sequence learning for activity recognition. Despite all advancements, existing systems for HAR face challenges in achieving high accuracy and for handling large datasets due to issues like variability in human actions, environmental complexity, and real-time processing needs. Further improvements in model architectures and training methodologies are needed to enhance HAR performance. We introduce feature extraction classification models for HAR to handle these issues with improved accuracy. This research's significant contributions are as follows:

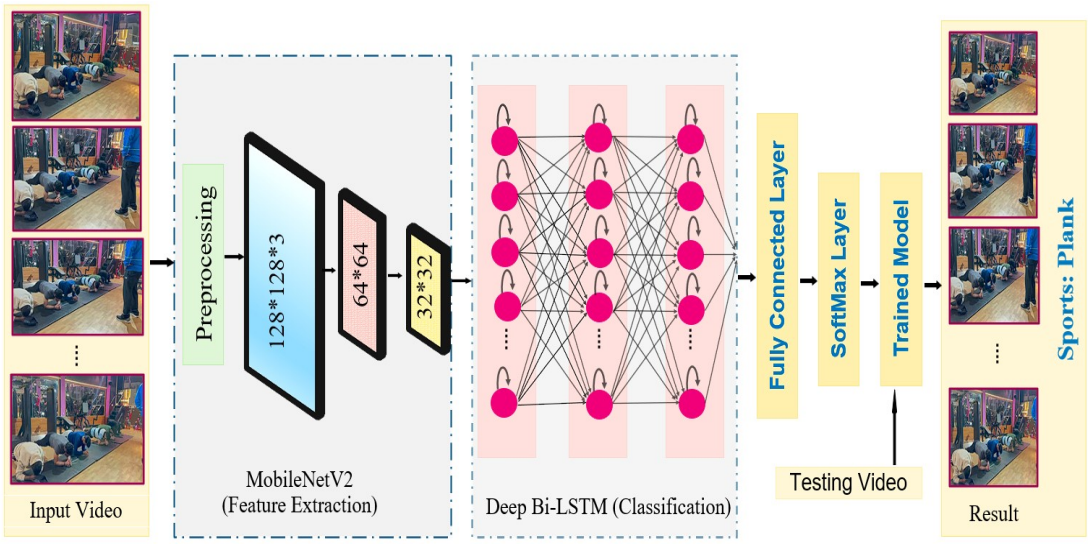


Fig. 7.1: Outline of proposed MV-DBiLSTM Classification Model

- We proposed the MV-DBiLSTM model for activity classification problems, integration of MobileNetV2, and Deep-BiLSTM to learn short-term dependencies as well as long-term features sequentially.

- We used MobileNetV2 for feature extraction, which significantly enriched inputs that improve the model's capability to recognize complex and dynamic patterns for accurate human activity recognition.
- Our method uses Deep Bi-LSTM to refine extracted features and capture complex temporal feature dependencies. This approach, processing data in both directions, i.e., forward and backward, helps to classify dynamic human activities while improving accuracy and recognition.
- The comprehensive experiments conducted to validate the efficacy of the proposed model on HAR benchmark datasets: HDMB51, UCF Sport, JHMDB, and one synthesized dataset. The results demonstrate its effectiveness and adaptability across diverse HAR scenarios.

7.2 Proposed Methodology: MV-DBiLSTM

The proposed methodology introduces a novel 'Deep MobileNetV2 Bi-directional Long Short Term Memory' termed MV-DBiLSTM classification model which refines and extracts the features from MobileNetV2, enhancing discrimination and capturing temporal dependencies for dynamic human actions. It efficiently handles diverse images, addressing the challenges of high-dimensional pixel data. The outline of our proposed MV-DBiLSTM classification model is illustrated in Figure 7.1, which integrates the functionalities of MobileNetV2 for feature extraction and Deep Bi-LSTM to capture spatial-temporal feature dependencies for dynamic activity recognition. MobileNetV2 efficiently handles dense connections and significantly boosts the model's performance. Its architecture permits capturing more effective feature extraction while maintaining computational efficiency, making it well-suited for processing complex visual data in HAR. Its strength lies in its ability to handle dense connections, enabling effective communication between pixel values and neurons. We optimized features using convolution, MaxPooling filter of 3*3, batch normalization, and dropout layers; then the output is flattened into a one-dimensional format, enabling efficient processing to follow through the dense layers. Relu and Softmax activation functions are employed to boost the computational efficiency of the model for classification predictions. Deep MV-Bi-LSTM refines hierarchical features extracted by MobileNetV2, enhancing discriminative power and capturing sequential dependencies in activity sequences. Bi-LSTM effectively handles long-term temporal dependencies in both directions, adapting dynamically to variable contexts for robust performance across diverse

situations. Configurable layer explorations and iterative fine-tuning further optimize the model's responsiveness to real-world discrepancies. Algorithm 1 summarizes the step-by-step methodology of our study, providing a clear representation of each phase involved in the proposed process. It includes pre-processing of video data, feature extraction using MobileNetV2, classification through the Bi-LSTM model, training and testing measures, and then final predictions.

In 1st step, we preprocess the input video data by converting each video sequentially into video frames. Further, these are arranged into frame format to prepare the data for further processing. We utilize a pre-trained model, i.e., MobileNetV2, in 2nd step, which extracts the meaningful features from the preprocesses data frame. MobileNetV2 is specifically selected due to its ability to handle computational complexity efficiently and use a multi-stage deep learning architecture, which enhances feature extraction capabilities. After feature extraction, we employ the Deep BiLSTM model in 3rd step for refining the extracted features and performing the classification task. This model helps capture temporal dependencies and improve human activity recognition accuracy. The model is trained using robust tuning techniques in 4th step. We iterate over multiple epochs during training and process the data in batches. Predictions from these batches are aggregated for training and testing datasets, and the resultants are utilized to compute the model's accuracy score. After model's get trained and validated, the outputs are displayed in accuracy and other performance metrics, demonstrating the model's efficiency in 5th step. In the final step, we aid real-time testing by allowing users to input video data. The model processes the input video and predicts the activity being performed. The output is accessible by providing a user-friendly interface for activity recognition.

7.2.1 MobileNetV2 for Feature Extraction

MobileNet [210] is a lightweight, efficient CNN model designed for mobile devices to optimize both size and performance. It uses depthwise separable convolutions to decrease computational complexity, which splits the convolution process into two stages: depthwise and pointwise. This makes MobileNet perfect for real-time image and video recognition with limited computational resources, like mobile and embedded systems. MobileNetV2 [211] is an innovative deep-learning architecture for efficient neural network operations and builds on the original MobileNet by introducing several innovations to enhance both performance and computational efficiency. Its flexibility supports users in regulating the size and dimensions of the network, making it adaptable to various computational requirements. The enhanced design of MobileNetV2, incor-

porating linear bottlenecks and skip connections, significantly improves computational efficiency while preserving high accuracy.

Our framework replaced the softmax unit with the Bi-LSTM model for predictions. In MobileNetV2 architecture, the centered residual blocks utilize depthwise and pointwise convolutional layers, with the pointwise layer functioning as a projection layer to reduce feature dimensions effectively. The architecture strikes stability amid computational efficiency and performance through linear bottlenecks and ReLU6 activations, ensuring efficient operations without compromising representational capability. The convolutional layer is unifying with a 3*3 layer, trailed by global average pooling for aggregation, and classify the output through dense layers. This combination optimizes the model for robust and efficient recognition tasks. MobileNetV2's adaptability, enabled by its customization width multiplier and resolution modifications, makes it highly recommended for smart devices with limited computational power for real-time applications. In our study, this model effectively extracted robust features from the dynamic human activity dataset, ensuring precise pattern recognition while maintaining efficiency. This combination of flexibility and performance utility validates resource-constrained environments requiring real-time processing.

7.2.2 Deep Bidirectional Long Short-Term Memory

Using only CNNs like MobileNetV2 for HAR may not fully capture temporal dependencies in activity sequences. To overcome this, we combined LSTM, intended to learn and retain information over time. RNNs are effective at capturing temporal information, but they struggle with long sequences due to exploding gradient problem. This challenge can destabilize training and hinder performance over extended durations. To address the limitations of traditional RNNs, we employed the specialized variant of LSTM, Deep Bi-LSTM [212], which is an extension of the Bi-LSTM model [213, 214]. Deep BiLSTM enhances the learning process by in view of both forward and backward temporal dependencies in sequences, making it more effective at capturing complex temporal relationships. This bidirectional approach improves the model's performance in recognizing dynamic human activities, addressing the shortcomings of standard LSTMs in handling long-term dependencies. Figure 7.2 depicts the outline of Deep Bi-LSTM classification model, highlighting its capability to effectively process and identify human activities. This model is the result of the traditional LSTM architecture, which relies on three primary gates, namely, the input, forget, and output gates, to adjust the flow of feature information.

Algorithm 1 Transfer-Learning-Feature-Based Human Activity Recognition

Input-V: List of input videos
    L: Ground truth labels for videos
Output:P: Predicted activity classes for input videos **Step 1: Preprocessing**
for each video $v \in V$ **do**
    $S \leftarrow$ VideoToImageConversion(v)
    Append S to $PreprocessedVideos$
end for
Step 2: Feature Extraction
Load MobileNetV2 model:$M \leftarrow$ loadModels$(MobileNetV2)$
Extract features: $VideoFeatures \leftarrow M(PreprocessedVideos)$
Step 3: Classification
Initialize Deep BiLSTM model: $ClassModels \leftarrow$ initialize$(DeepBiLSTM)$
Pass extracted features: $PredictedClass \leftarrow ClassModels(VideoFeatures)$
Step 4: Training
for NumEpochs **do**
    Divide $VideoFeatures$ to training batches: B_{train}
    for each batch b **do**
        Predict class: $\hat{y} \leftarrow ClassModels(b)$
        Compute loss: $Loss \leftarrow$ Criterion$(\hat{y}, L)$
        Update model: $Loss$.backward(), Optimizer.Step()
    end for
end for
Step 5: Evaluation
Divide $VideoFeatures$ into testing batches: B_{test}
for each batch b **do**
    Predict class: $y_{test} \leftarrow ClassModels(b)$
    Evaluate performance: $Performance \leftarrow$ PerformanceMatrix(y_{test}, L)
end for
Step 6: Display Results
Print accuracy & performance metrics: print$("PerformanceMetrics : ", Performance)$ **Step 7: Real-Time Testing**
Record or input real-time video: $TestVideo \leftarrow$ Recorded$(Video)$
Convert video to frames: $TestFrames \leftarrow$ VideoToImageConversion$(TestVideo)$
Predict activity: $PredictedActivity \leftarrow ClassModels(TestFrames)$
Display result: print$("PredictedActivity : ", PredictedActivity)$
end

7.2.2.1 Bi-LSTM

Bi-LSTM overcomes the limitation of traditional LSTM by processing both past and future sequences. It uses two hidden layers: the forward pass layer and the backward

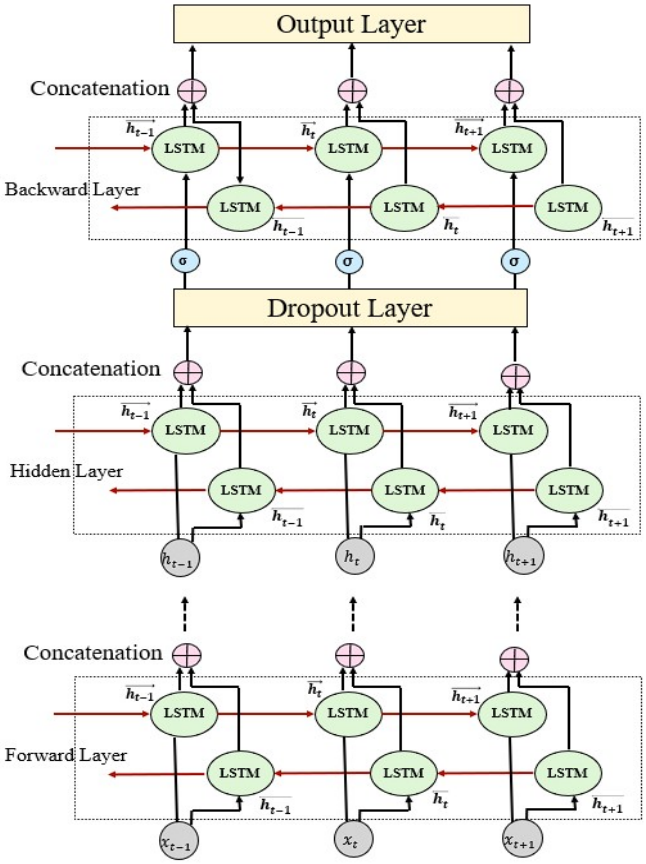


Fig. 7.2: Framework of Deep Bi-directional LSTM

pass layer. These two layers combine their output to generate a final prediction. The forward hidden sequence $\overrightarrow{h}$ considers previous inputs, while the backward hidden sequence $\overleftarrow{h}$ looks at future inputs. This bidirectional approach helps capture more comprehensive temporal dependencies in sequential data. The final output is derived by combining both forward and backward states y for grasping the long-range temporal feature from frames time $t = 1$ to T, as follows:

$$\overrightarrow{h_t} = \sigma\left(\omega_{x\overrightarrow{h}} x_t + \omega_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}\right) \tag{7.1}$$

$$\overleftarrow{h_t} = \sigma\left(\omega_{x\overleftarrow{h}} x_t + \omega_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}\right) \tag{7.2}$$

$$y_t = \omega_{\overrightarrow{h}y} \overrightarrow{h}_t + \omega_{\overleftarrow{h}y} \overleftarrow{h}_t + b_y \tag{7.3}$$

where σ represents the activation function, ω is the weight matrices. b indicates the bias terms. x_t, $\overrightarrow{h_t}$, $\overleftarrow{h_t}$ indicates input and hidden states.

7.2.2.2 Proposed Deep Bi-LSTM Sequence Model

The traditional Bi-LSTM network operates an input sequences in both directions- forward and backward, which makes it fit for tasks that need context from both past and future elements. We proposed a deep Bi-LSTM network that replaces the LSTM cells to learn the temporal dynamics from multiple data streams. The Deep Bi-LSTM processes distinct input sequences via hidden layer h_n for each input vector x_n. Unlike a standard BiLSTM, which utilizes only two hidden layers, the proposed Deep BiLSTM incorporates multiple hidden layers. The weight matrix ω is employed to transform inputs from x-space to h-space. A single LSTM model is not optimal for combining inputs from multiple independent sequences, as it struggles to establish a unified subspace to effectively map heterogeneous input representations. Therefore, for each n^{th} stream $(1 \leq n \leq N)$, the Deep BiLSTM takes the input sequence $(n^n = \{x_1^n, x_2^n, x_3^n, \ldots, x_T^n\})$, and computes the hidden sequence $(h^n = \{h_1^n, h_2^n, h_3^n, \ldots, h_T^n\})$, to predict the output y by iterating $t = 1$ to T. Their mathematical equations for the hidden states and output are as follows:

$$h_t^N = \sigma\left(\omega_{xh^N}^N X_t^N + \omega_{hh^N}^N h_{t-1}^N + b_h^N\right), \tag{7.4}$$

$$h_t^2 = \sigma\left(\omega_{xh^2}^2 X_t^2 + \omega_{hh^2}^2 h_{t-1}^2 + b_h^2\right), \tag{7.5}$$

$$h_t^1 = \sigma\left(\omega_{xh^1}^1 X_t^1 + \omega_{hh^1}^1 h_{t-1}^1 + b_h^1\right), \tag{7.6}$$

$$y_t = \sum_{n=1}^{N} \omega_{h^n y} h_t^n + b_y. \tag{7.7}$$

where N indicates a number of streams, which is used to capture the complex spatiotemporal patterns of videos. X_t^n represents input features from the nth stream at time t. h_t^n is the hidden state of the nth stream at time t. σ acts as an activation function applied to the weighted sum. $\omega_{xh^n}^n$ is the weight matrices for the input-to-hidden and hidden-to-hidden layers, respectively. b_h^n is the bias weight assign for the nth stream. y_t: The final output at time t, which aggregates the contributions from all streams.

7.3 Dataset

The datasets used in this study for our experiments are HMDB51, JHMB, UCF Sports, and one synthesized dataset. While numerous datasets are available, we selected them because they present more significant challenges, requiring advanced techniques to enhance recognition performance and effectively address their inherent complexities.

7.3.1 HMDB51 Dataset

The dataset [151] comprises 6,849 clips, categorized in 51 distinct classes; each action class contains a minimum of 101 clips. These action categories are further grouped into broader types such as jump, smile, eat, kiss, etc. The original evaluation employs on three different split training/testing. Every action class includes 70 video clips for training and 30 for testing.

7.3.2 Joint Annotated Human Motion DataBase (JHMDB)

The dataset [215] consists of 960 video sequences comprising 21 different classes of activities such as catch, clamp, hair brush, baseball, clap, swing, gunshot fire, etc. The dataset have video and annotation for puppet flow per frame, puppet mask per frame, joint positions per frame, action label per clip and meta label per clip.

7.3.3 UCF Sports Dataset

The dataset [216][217] has 150 video sequence clips at 720*480 resolution, featuring sports activities like horse riding, golf, swings, driving, jumping, skateboarding, weightlifting, etc. Sourced from outlets such as BBC and ESPN, the videos showcase authentic actions from various perspectives and scenes.

7.3.4 Synthetic Dataset

We developed a synthesized human action dataset in the Delhi Technological University campus and for indoor in the Software Engineering laboratory; till now, out of 250, only 150 video clips have been synthesized. The video clips are recorded through a camera of 12 Megapixels, 1.7µm with PDAF dual pixel and sensor-shift OIS at a resolution of 480*640. Currently, it contains 6 activity classes of sports, i.e., Indoor Sports, Outdoor Sports, Exercise, Yoga, Fight, and Dancing. Each class further has many categories of the same type such as, plank, rafting, push-ups, barbell squat, running, basketball, boxing, table tennis, badminton, and balling etc. We have chosen 105 video clips as a random sample for training purposes, and the rest for testing.

7.4 Experimental Results and Discussion

We validated the proposed model through extensive experiments on challenging HAR datasets. The outcomes demonstrated that the proposed model performs better than

SOTA models in accuracy, with additional metrics also being reported. In the subsequent sections, we summarize the training process and quantitative and qualitative comparisons. Additionally, we conducted detailed ablation training on each dataset to evaluate the model's strengths and weaknesses comprehensively.

7.4.1 Implementation Setup and Evaluation Metrics

The experiment is executed in Python using Keras & TensorFlow. The experiments are implemented over an AMD Ryzen7 NVIDIA GEFORCE RTX GPU, & RAM 132 GB. The process utilized ConvNet to extract CNN features, along with MobileNetV2, and employed neural network dependencies for implementing Deep BiLSTM. The extracted features are concatenated with a temporal dimension and input into Deep BiLSTM for sequence learning. Subsequently, a shared LSTM layer with 16 units is employed for feature fusion and to remove irrelevant noisy information, ensuring dimensional consistency. Further, the filtered features then passed to a fully connected layer. Overall, the model is trained on 150 epochs having batch size of 64 with learning rate 0.0001. The proposed method's performance is calculated over metrics such as precision, recall, F1-score, and accuracy. The mathematical formulations are:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7.8}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{7.9}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7.10}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7.11}$$

where TP, TN, FP, and FN indicate the totality of true positives, true negatives, false positives, and false negatives samples, respectively.

7.4.2 Experiments on HMDB51 Dataset

The proposed model is evaluated against SOTA methods on the HMDB51 dataset, as shown in Table 7.1, which includes challenging action recognition tasks. The outcomes indicate that the proposed model MV-Deep Bi-LSTM achieves the highest accuracy of 77.5242%, surpassing all other techniques.

Table 7.1: Comparative Analysis with SOTA Methods on HMDB51 Dataset

Techniques	Accuracy (%)
DD-Net [218]	68.01
Dilated CNN + BiLSTM + RB [164]	72.64
Two-stream LSTM [219]	73.20
STAD-ConvBi-LSTM [61]	69.54
MV DBi-LSTM (Proposed)	**77.52**

7.4.3 Experiments on UCF Sports Dataset

Table 7.2 shows the comparison of the proposed model against SOTA methods on UCF Sports Dataset. The proposed method MV-DBiLSTM achieves the highest accuracy of 94.0052%, outperforming all other approaches.

Table 7.2: Comparative Analysis with SOTA Methods on UCF Sports Dataset

Techniques	Accuracy (%)
Dilated CNN + BiLSTM + RB [164]	89.64
Two-stream LSTM [219]	89.20
STAD-ConvBi-LSTM [61]	92.12
MV DBi-LSTM (Proposed)	**94.00**

7.4.4 Experiments on JHMDB Dataset

Table 7.3 showcases a comparative analysis of the proposed model against SOTA methods on JHMDB Dataset. Our proposed method achieves the highest accuracy of 96.785%, significantly outperforming other approaches.

Table 7.3: Comparative Analysis with SOTA Methods on JHMDB Sports Dataset

Techniques	Accuracy (%)
Hybrid Deep Neural Network [220]	88.28
DD-Net [218]	78.00
Two-stream LSTM [219]	92.70
STAD-ConvBi-LSTM [61]	90.01
MV DBi-LSTM (Proposed)	**96.78**

7.4.5 Experiments on Synthesized Dataset

Table 7.4 showcases a comparative analysis of the proposed model against SOTA methods on Synthesized Dataset. The proposed method achieves the highest accuracy of 96.79%, significantly outperforming other approaches.

Table 7.4: Comparative Analysis with SOTA Methods on Synthesized Dataset

Techniques	**Accuracy (%)**
Hybrid Deep Neural Network [221]	72.65
Two-stream LSTM [219]	88.24
STAD-ConvBi-LSTM [61]	92.89
MV DBi-LSTM (Proposed)	**96.82**

7.5 Summary

This work presents MV-DBiLSTM Classification Model that combines MobileNetV2 for spatial feature extraction and bi-directional LSTM for capturing temporal dependencies in HAR from video data. The model integrates MobileNetV2, a lightweight CNN, to extract meaningful spatial features from individual video frames. These features are then passed to a Bi-LSTM network, which effectively captures temporal dependencies in both forward and backward directions, enabling the model to learn contextual patterns across the full video sequence. A softmax activation function is employed for final classification, enhancing the model's ability to assign accurate class probabilities. The proposed approach was rigorously tested on three public benchmark datasets as well as a synthesized dataset. These results highlight the robustness, generalizability, and efficiency of the MV-DBiLSTM model in handling complex and dynamic human actions.

Chapter 8

Conclusion, Future Scope and Social Impact

8.1 Conclusion

This thesis presented a comprehensive exploration of AR, focusing on the development of robust, adaptive, & intelligent ML-DL based frameworks capable of handling a wide range of challenges in real-world environments. Throughout this work, several novel models were proposed to improve recognition performance under varying conditions such as occlusion, dynamic motion, low-light visibility, and complex temporal dependencies. The research began with a systematic review of SOTA for these systems, highlighting the taxonomy of techniques, limitations in existing datasets, and evolution of ML-DL architectures. This foundational analysis identified research gaps, particularly in model generalizability, real-time applicability, and environmental adaptability. The research presented in this thesis was guided by the four objectives outlined in Section 1.8. The following points revisit each objective and summarize the key outcomes achieved.

- **Objective 1** focused on designing an effective framework for accurately detecting and classifying human activities across diverse benchmark datasets. This objective was successfully achieved through the development of multiple DL-based architectures such as ConvST-LSTM-Net and STAD-ConvBi-LSTM. Experimental evaluation on multiple public datasets demonstrated strong classification accuracy, validating the effectiveness of the proposed solutions.
- **Objective 2** aimed to conduct a comprehensive evaluation of existing literature & synthesized datasets to understand model behavior under controlled and perturbed environments. This objective was accomplished through an extensive SLR and through empirical analysis across light, low-light, occlusion, and multi-view conditions. The results provided insights into model stability, strengths,

and limitations under different operational scenarios.

- **Objective 3** involved developing a robust multi-level feature fusion approach capable of handling partial occlusion. This objective was realized through the MSPAST-GCN architecture, which integrates spatial–temporal graph features with attention-based fusion. Experiments on synthesized occlusion datasets confirmed that the model effectively retains discriminative features even when several key joints or body segments are occluded.

- **Objective 4** aimed to design adaptive and dynamic models capable of performing reliably under variation in illumination, background clutter, occlusion, and motion complexity. This was achieved through the integration of image enhancement modules, illumination correction, and multi-stream processing pipelines. This demonstrates how the combination of enhancement techniques and DETR-based vision transformers significantly improves recognition performance in challenging environments.

Collectively, the thesis meets all four research objectives, demonstrating a coherent progression from initial problem formulation to model development, experimentation, analysis, and domain-specific conclusions. The results confirm that the proposed methodologies are both technically sound and practically applicable to real-world HAR scenarios.

8.2 Future Scope

The methods and models proposed in this thesis mark a significant impact in the field of AR using ML-DL, particularly in capturing complex spatial-temporal dependencies from human motion data. The integration of the proposed HAR models with modern LLM-based vision–language systems can create a powerful synergy that directly contributes to social good. LLMs such as GPT-4V, CLIP, PaLI, and VideoGPT offer high-level semantic reasoning, contextual understanding, natural-language explanation, and decision support [222, 223, 224, 225]. When combined, HAR models can provide reliable and structured visual representations that LLMs can interpret, explain, and communicate in human-understandable formats. This mutual reinforcement opens up impactful opportunities for social good: real-time fall detection and safety monitoring for the elderly, precise behavioral tracking in healthcare, improved surveillance for public safety while maintaining transparency and ethical oversight, enhanced assistance for visually impaired individuals using natural-language activity descriptions

and intelligent disaster-response systems that can detect risky human behaviors and generate context-aware warnings. In essence, the combination of robust HAR feature extraction with the interpretability and reasoning capabilities of LLMs can lead to next-generation human-centered AI systems that are not only technologically advanced but also socially responsible, transparent, and accessible. The following key directions are identified for future exploration:

- While the current models demonstrate strong recognition performance, they are not yet optimized for real-time deployment on resource-constrained devices such as surveillance cameras, mobile phones, or wearable sensors. Future work will focus on designing lightweight, low-latency architectures that maintain high accuracy while enabling seamless integration into real-world, real-time applications. Techniques such as model pruning, quantization, and hardware-specific optimization (e.g., for GPUs, TPUs) will be explored.

- This thesis primarily utilized visual data i.e., RGB frames, skeleton keypoints for AR. Future research will incorporate multimodal sensor fusion, including audio signals, depth maps, IMU data, and thermal imaging, to improve model robustness in complex scenarios such as low visibility, noisy backgrounds, or occluded views. Combining modalities can enhance recognition accuracy and reliability in real-world conditions.

- A major challenge in training accurate AR models is the need for large amounts of labeled data. Future work will investigate self-supervised, semi-supervised, and few-shot learning techniques, enabling models to learn effective representations with minimal human annotation. This will significantly improve the scalability and adaptability of these systems across different domains and datasets.

- Current datasets often consist of scripted, clean activity sequences that do not fully represent real-world variability. Future work will aim to develop diverse and realistic HAR datasets with multiview setups, occlusions, spontaneous behaviors, and varying lighting or environmental conditions. These datasets will be essential for benchmarking and improving the generalizability of next-generation activity recognition systems.

- Building on the current framework, future research will extend the application of activity recognition to other critical domains such as workplace safety, sports performance analysis, elderly care, and emergency response systems. Quantitative studies will assess the impact of AR systems in real-world applications,

aiming to improve operational efficiency, safety, and decision-making in both public and private sectors.

- Another important future direction is the development of lightweight, on-device multimodal HAR–LLM systems optimized for edge and mobile environments. Current LLMs are computationally intensive, making them unsuitable for real-time deployment in surveillance cameras, IoT networks, elderly-care monitoring systems, or smart homes. Future research can explore model compression, distillation, and hardware-aware optimization to integrate the proposed HAR architectures with compact vision–language models capable of running efficiently on edge processors while maintaining high recognition accuracy and interpretability.

- Finally, future work will explore unconventional computing techniques and emerging AI paradigms that may enhance the capabilities of AR systems. This includes the potential use of quantum computing, neuromorphic hardware, or novel neural architectures to meet the growing demands of complex, real-time, and context-aware activity recognition.

By addressing these directions, future research will not only extend the capabilities of current activity recognition systems but also pave the way for smarter, more responsive, and socially impactful technologies across a wide range of sectors.

8.3 Social Impact

The outcomes of this research carry significant and far-reaching implications across diverse sectors that depend on intelligent behavior analysis and human activity understanding. From a societal perspective, the proposed activity recognition frameworks contribute meaningfully to enhancing public safety, particularly in critical and high-risk environments such as airports, metro stations, public gatherings, and transportation hubs. By enabling early detection or recognition of abnormal or suspicious behaviors such as loitering, aggression, or unauthorized access. These systems help prevent potential security threats, reduce the burden on human operators, and facilitate faster, more effective interventions. This directly supports the development of safer public spaces and the advancement of smart city infrastructure. Beyond public safety, these systems also promote comfort, efficiency, and inclusivity across multiple domains.

In healthcare, AR systems support elderly care, patient monitoring, and rehabilitation

by recognizing critical events such as falls or prolonged inactivity in real time. This improves patient safety, encourages independent living, and helps reduce long-term healthcare costs through preventive monitoring in hospitals, clinics, and home-care settings. In smart environments such as automated homes, intelligent offices, malls, and adaptive learning spaces, these systems enable context-aware automation, adjusting lighting, temperature, and security in response to detected human activity, thus enhancing both energy efficiency and user comfort. In education, gesture- and motion-based interfaces powered by AR make digital learning more interactive and personalized, especially in virtual classrooms and e-learning platforms, boosting student engagement and accessibility. In entertainment, these systems create responsive and immersive experiences in AR/VR and gaming environments by adapting in real time to user movements and behavior. AR also plays a crucial role in assistive technologies, promoting social inclusivity by enabling individuals with physical or cognitive disabilities to interact more independently with their surroundings. These systems support voice-free control, motion-triggered actions, and adaptive interfaces, enhancing accessibility in both digital and physical environments. In workplace safety, AR systems monitor hazardous environments, detect unsafe behaviors, and issue alerts to help prevent accidents, making industrial operations safer and more compliant. In sports and training, these systems enable detailed analysis of posture, movement, and technique, helping coaches and athletes improve performance and reduce injury risks through data-driven insights. Within transportation systems, AR is used to monitor passenger behavior, detect overcrowding, and identify emergencies, particularly in public transit, thereby improving both service reliability and rider safety. In the context of disaster and emergency response, AR technologies aid in real-time monitoring and tracking, helping locate survivors, assess movement patterns, and coordinate effective rescue efforts.

Collectively, these applications demonstrate the transformative social potential of activity recognition technologies to improve safety, efficiency, accessibility, and quality of life in today's increasingly connected and intelligent digital world.

Bibliography

[1] R. Singh, A. K. S. Kushwaha, R. Srivastava *et al.*, "Recent trends in human activity recognition–a comparative study," *Cognitive Systems Research*, vol. 77, pp. 30–44, 2023.

[2] A. Wang, G. Chen, J. Yang, S. Zhao, and C.-Y. Chang, "A comparative study on human activity recognition using inertial sensors in a smartphone," *IEEE Sensors Journal*, vol. 16, no. 11, pp. 4566–4578, 2016.

[3] F. Demrozi, G. Pravadelli, A. Bihorac, and P. Rashidi, "Human activity recognition using inertial, physiological and environmental sensors: A comprehensive survey," *IEEE Access*, vol. 8, pp. 210 816–210 836, 2020.

[4] C. Dhiman and D. K. Vishwakarma, "A review of state-of-the-art techniques for abnormal human activity recognition," *Engineering Applications of Artificial Intelligence*, vol. 77, pp. 21–45, 2019.

[5] J. Wang, L. Hu, Y. Chen, S. Hao, and H. Pan, "Deep learning for sensor-based activity recognition: a survey," *Pattern Recognition Letters*, pp. 3–11, 2019.

[6] S. R. Ke, H. L. N. Thuc, Y. J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A review on video-based human activity recognition," *Computers*, vol. 2, no. 2, pp. 88–131, 2013.

[7] A. Jalal, M. Uddin, and T.-S. Kim, "Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 3, pp. 863–871, 2012.

[8] J. Perš, V. Sulić, M. Kristan, M. Perše, K. Polanec, and S. Kovačič, "Histograms of optical flow for efficient representation of body motion," *Pattern Recognition Letters*, vol. 31, no. 11, pp. 1369–1376, 2010.

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, pp. 886–893 vol. 1.

[10] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *European Conference on Computer Vision (ECCV)*, ser. Lecture Notes in Computer Science. Springer, 2006, pp. 428–441.

[11] N. Cristianini and E. Ricci, *Support Vector Machines*. Boston, MA: Springer US, 2008, pp. 928–932.

[12] R. M. Schmidt, "Recurrent neural networks (rnns): A gentle introduction and overview," 2019.

[13] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.

[14] M. M. Rahman, Y. Watanobe, and K. Nakamura, "A bidirectional lstm language model for code evaluation and repair," *Symmetry*, vol. 13, no. 2, p. 247, 2021.

[15] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (gru) neural networks," in *2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2017, pp. 1597–1600.

[16] T. Ojala, M. Pietikäinen, and D. Harwood, "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," in *Proceedings of the 12th IAPR International Conference on Pattern Recognition (ICPR)*, vol. 1. IEEE, 1994, pp. 582–585.

[17] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *Computer Vision – ECCV 2006*, ser. Lecture Notes in Computer Science, A. Leonardis, H. Bischof, and A. Pinz, Eds. Springer, Berlin, Heidelberg, 2006, vol. 3951, pp. 404–417.

[18] I. U. Khan, S. Afzal, and J.-W. Lee, "Human activity recognition via hybrid deep learning based model," *Sensors*, vol. 22, no. 1, p. 323, 2022.

[19] W. Zhang, M. Tomizuka, and N. Byl, "A wireless human motion monitoring system based on joint angle sensors and smart shoes," in *Dynamic Systems and Control Conference*, vol. 46209. American Society of Mechanical Engineers, 2014, p. V003T46A002.

[20] A. Sanchez-Caballero, S. de Lopez-Diz, D. Fuentes-Jimenez, C. Losada-Gutierrez, M. Marron-Romera, D. Casillas-Perez *et al.*, "3dfcnn: Real-time action recognition using 3d deep neural networks with raw depth information," *Multimedia Tools and Applications*, pp. 1–25, 2022.

[21] J. Chen, Y. Sun, and S. Sun, "Improving human activity recognition performance by data fusion and feature engineering," *Sensors*, vol. 21, no. 3, pp. 1–23, 2021.

[22] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *European Conference on Computer Vision (ECCV)*, 2020.

[23] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," 2016.

[24] R. Zhu *et al.*, "Efficient human activity recognition solving the confusing activities via deep ensemble learning," *IEEE Access*, vol. 7, pp. 75 490–75 499, 2019.

[25] M. Karim, S. Khalid, A. Aleryani, J. Khan, I. Ullah, and Z. Ali, "Human action recognition systems: A review of the trends and state-of-the-art," *IEEE Access*, vol. 12, pp. 12 345–12 356, Mar 2024.

[26] M. A. R. Ahad, A. D. Antar, and O. Shahid, "Vision-based action understanding for assistive healthcare: A short review." in *CVPR workshops*, vol. 2, 2019.

[27] W. Chen, C. Yu, C. Tu, Z. Lyu, J. Tang, S. Ou, Y. Fu, and Z. Xue, "A survey on hand pose estimation with wearable sensors and computer-vision-based methods," *Sensors*, vol. 20, no. 4, p. 1074, 2020.

[28] H. M. Do, K. C. Welch, and W. Sheng, "Soham: A sound-based human activity monitoring framework for home service robots," *IEEE Transactions on Automation Science and Engineering*, vol. 19, no. 3, pp. 2369–2383, 2021.

[29] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. Béjar, D. D. Yuh *et al.*, "Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling," in *MICCAI workshop: M2cai*, vol. 3, no. 2014, 2014, p. 3.

[30] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *International Journal of Computer Vision*, vol. 130, no. 5, pp. 1366–1401, 2022.

[31] G. Lan, Y. Wu, F. Hu, and Q. Hao, "Vision-based human pose estimation via deep learning: A survey," *IEEE Transactions on Human-Machine Systems*, vol. 53, no. 1, pp. 253–268, 2023.

[32] Y. Karayaneva, S. Sharifzadeh, Y. Jing, K. Chetty, and B. Tan, "Sparse feature extraction for activity detection using low-resolution ir streams," in *2019 18th IEEE International Conference on Machine Learning and Applications*. IEEE, 2019, pp. 1837–1843.

[33] J.-K. Kim, K. Lee, and S. G. Hong, "Detection of important features and comparison of datasets for fall detection based on wrist-wearable devices," *Expert Systems with Applications*, vol. 234, p. Article 121034, 2023.

[34] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 156–165.

[35] N. Gupta *et al.*, "Human activity recognition in artificial intelligence framework: A narrative review," *Artificial Intelligence Review*, vol. 55, no. 6, pp. 4755–4808, 2022.

[36] C. Xu, D. Chai, J. He, X. Zhang, and S. Duan, "Innohar: A deep neural network for complex human activity recognition," *IEEE Access*, vol. 7, pp. 9893–9902, 2019.

[37] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational Intelligence and Neuroscience*, vol. 2018, p. 7068349, Feb 2018.

[38] M. G. Morshed, T. Sultana, A. Alam, and Y.-K. Lee, "Human action recognition: A taxonomy-based survey, updates, and opportunities," *Sensors*, vol. 23, no. 4, p. 2182, Feb 2023.

[39] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognition Letters*, vol. 119, pp. 3–11, Mar 2019.

[40] M. Al-Faris, J. Chiverton, D. Ndzi, and A. I. Ahmed, "A review on computer vision-based methods for human action recognition," *Journal of imaging*, vol. 6, no. 6, p. 46, 2020.

[41] Y. Bai, D. Zhou, S. Zhang, J. Wang, E. Ding, Y. Guan, Y. Long, and J. Wang, "Action quality assessment with temporal parsing transformer," in *European conference on computer vision*. Springer, 2022, pp. 422–438.

[42] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[43] Q. Lei, J.-X. Du, H.-B. Zhang, S. Ye, and D.-S. Chen, "A survey of vision-based human action evaluation methods," *Sensors*, vol. 19, no. 19, p. 4129, 2019.

[44] R. Hou, C. Chen, and M. Shah, "Tube convolutional neural network (t-cnn) for action detection in videos," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5822–5831.

[45] A. Kanade, M. Sharma, and M. Muniyandi, "Tele-evalnet: A low-cost, teleconsultation system for home-based rehabilitation of stroke survivors using multiscale cnn-convlstm architecture," in *LNCS: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13806, 2023, pp. 738–750.

[46] X. Peng and C. Schmid, "Multi-region two-stream r-cnn for action detection," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part IV*. Springer, 2016, pp. 744–759.

[47] A. Dadashzadeh, S. Duan, A. Whone, and M. Mirmehdi, "Pecop: Parameter efficient continual pretraining for action quality assessment," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 42–52.

[48] Z. Gan, L. Jin, Y. Cheng, Y. Cheng, Y. Teng, Z. Li, Y. Li, W. Yang, Z. Zhu, J. Xing *et al.*, "Skatingverse: A large-scale benchmark for comprehensive evaluation on human action understanding," *IET Computer Vision*, vol. 18, no. 7, pp. 888–906, 2024.

[49] T. He, H. Liu, Y. Li, X. Ma, C. Zhong, Y. Zhang, and W. Lin, "Collaborative weakly supervised video correlation learning for procedure-aware instructional video analysis," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 3, 2024, pp. 2112–2120.

[50] X. Ke, H. Xu, X. Lin, and W. Guo, "Two-path target-aware contrastive regression for action quality assessment," *Information Sciences*, vol. 664, 2024.

[51] S. Sardari, S. Sharifzadeh, A. Daneshkhah, S. W. Loke, V. Palade, M. J. Duncan *et al.*, "Lightpra: A lightweight temporal convolutional network for automatic physical rehabilitation exercise assessment," *Computers in Biology and Medicine*, vol. 173, 2024.

[52] B. X. Yu, Y. Liu, K. C. Chan, and C. W. Chen, "Egcn++: A new fusion strategy for ensemble learning in skeleton-based rehabilitation exercise assessment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–16, 2024.

[53] K. Roditakis, A. Makris, and A. Argyros, "Towards improved and interpretable action quality assessment with self-supervised alignment," in *ACM International Conference Proceeding Series*, 2021, pp. 507–513.

[54] V. Rani, S. T. Nabi, M. Kumar, A. Mittal, and K. Kumar, "Self-supervised learning: A succinct review," *Archives of Computational Methods in Engineering*, vol. 30, no. 4, pp. 2761–2775, 2023.

[55] I. Molenaar, S. de Mooij, R. Azevedo, M. Bannert, S. Järvelä, and D. Gašević, "Measuring self-regulated learning and the role of ai: Five years of research using multimodal multichannel data," *Computers in Human Behavior*, vol. 139, p. 107540, 2023.

[56] M. C. Schiappa, Y. S. Rawat, and M. Shah, "Self-supervised learning for videos: A survey," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–37, 2023.

[57] Y. Fang, Z. Luo, F. Huang, Z. Wang, D. Li, and X. Hua, "Developing a mixed reality-based game for post-stroke motor rehabilitation: Combining training and assessment," in *2023 9th International Conference on Virtual Reality (ICVR)*. IEEE, 2023, pp. 393–399.

[58] M. Nekoui, F. Cruz, and L. Cheng, "Eagle-eye: Extreme-pose action grader using detail bird's-eye view," in *Proceedings - 2021 IEEE Winter Conference on Applications of Computer Vision*, 2021, pp. 394–402.

[59] A. Ullah, K. Muhammad, W. Ding, V. Palade, I. U. Haq, and S. W. Baik, "Efficient activity recognition using lightweight cnn and ds-gru network for surveillance applications," *Applied Soft Computing*, vol. 103, p. 107102, 2021.

[60] D. Kurchaniya and S. Kumar, "Two stream deep neural network based framework to detect abnormal human activities," *Journal of Electronic Imaging*, vol. 32, no. 4, pp. 043 021–043 021, 2023.

[61] A. Sharma and R. Singh, "Convst-lstm-net: convolutional spatiotemporal lstm networks for skeleton-based human action recognition," *International Journal of Multimedia Information Retrieval*, vol. 12, no. 2, p. 34, 2023.

[62] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, "A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions," *Knowledge-Based Systems*, vol. 223, p. 106970, 2021.

[63] H. F. Nweke, Y. W. Teh, G. Mujtaba, and M. A. Al-Garadi, "Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions," *Information Fusion*, vol. 46, pp. 147–170, 2019.

[64] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tang, "Rgb-d-based action recognition datasets: A survey," *Pattern Recognition*, vol. 60, pp. 86–105, 2016.

[65] L. Martínez-Villaseñor, H. Ponce, J. Brieva, E. Moya-Albor, J. Núñez-Martínez, and C. Peñafort-Asturiano, "Up-fall detection dataset: A multimodal approach," *Sensors*, vol. 19, no. 9, p. 1988, 2019.

[66] S. Zhang, Z. Wei, J. Nie, L. Huang, S. Wang, and Z. Li, "A review on human activity recognition using vision-based method," *Journal of healthcare engineering*, vol. 2017, no. 1, p. 3090343, 2017.

[67] P. N. Deelaka, D. Y. Silva, S. Wickramanayake, D. Meedeniya, and S. Rasnayaka, "Sez-harn: Self-explainable zero-shot human activity recognition network," September 2023, under review at AAAI Conference.

[68] V. Estevam *et al.*, "Zero-shot action recognition in videos: A survey," *Neurocomputing*, vol. 453, pp. 112–130, 2021.

[69] K. Huang, L. Miralles-Pechuán, and S. McKeever, "Enhancing zero-shot action recognition in videos by combining gans with text and images," *SN Computer Science*, vol. 4, p. 375, 2023.

[70] E. Shabaninia, H. Nezamabadi-pour, and F. Shafizadegan, "Transformers in action recognition: A review on temporal modeling," *arXiv preprint arXiv:2302.01921*, 2022.

[71] M. M. Islam, S. Nooruddin, F. Karray, and G. Muhammad, "Human activity recognition using tools of convolutional neural networks: A state of the art review, data sets, challenges, and future prospects," *Computers in biology and medicine*, vol. 149, p. 106060, 2022.

[72] A. Ahmad, M. F. Zuhairi, S. Musa, F. Alanazi, A. Namoun, and A. Alrehaili, "A brief review of graph convolutional neural network based learning for classifying remote sensing images," *Procedia Computer Science*, vol. 191, pp. 540–545, 2021.

[73] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, 2014, pp. 568–576.

[74] G. Yao, T. Lei, and J. Zhong, "A review of convolutional-neural-network-based action recognition," *Pattern Recognition Letters*, vol. 118, pp. 14–22, 2019.

[75] G. Sreenu and M. A. Saleem Durai, "Intelligent video surveillance: a review through deep learning techniques for crowd analysis," *Journal of Big Data*, vol. 6, no. 1, p. 48, 2019.

[76] C. Yin, X. Miao, J. Chen, H. Jiang, D. Chen, Y. Tong, and S. Zheng, "Human activity recognition with low-resolution infrared array sensor using semi-supervised cross-domain neural networks for indoor environment," *arXiv preprint arXiv:2403.02632*, 2024.

[77] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, and Y. Liu, "Deep learning for sensor-based human activity recognition: Overview, challenges and opportunities," *arXiv preprint arXiv:2001.07416*, 2020.

[78] B. Nguyen, Y. Coelho, T. Bastos *et al.*, "Trends in human activity recognition with focus on machine learning and power requirements," *Machine Learning with Applications*, vol. 5, p. 100072, 2021.

[79] F. Duan, T. Zhu, J. Wang, L. Chen, H. Ning, and Y. Wan, "A multi-task deep learning approach for sensor-based human activity recognition and segmentation," *arXiv preprint arXiv:2303.11100*, 2023.

[80] L. M. Dang, K. Min, H. Wang, M. J. Piran, C. H. Lee, and H. Moon, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognition*, vol. 108, p. 107561, 2020.

[81] D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, "Vision-based human activity recognition: a survey," *Multimedia Tools and Applications*, vol. 79, pp. 30 509–30 555, 2020.

[82] L. Wang, D. Q. Huynh, and P. Koniusz, "A comparative review of recent kinect-based action recognition algorithms," *arXiv preprint arXiv:1906.09955*, 2019.

[83] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "Rgb-d-based human motion recognition with deep learning: A survey," *Computer Vision and Image Understanding*, vol. 171, pp. 118–139, 2018.

[84] A. Chakrabarti, N. Dey, F. Shi, and S. K. Saha, "Vision and inertial sensing fusion for human action recognition: A review," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 6, pp. 619–628, 2021.

[85] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu, "Human action recognition from various data modalities: A review," *arXiv preprint arXiv:2012.11866*, 2020.

[86] S. Majumder and N. Kehtarnavaz, "Vision and inertial sensing fusion for human action recognition: A review," *IEEE Sensors Journal*, vol. 21, no. 3, pp. 2454–2467, 2021.

[87] J. Li, H. Cui, T. Guo, Q. Hu, and Y. Shen, "Efficient fitness action analysis based on spatio-temporal feature encoding," in *2020 IEEE International Conference on Multimedia & Expo Workshops*. IEEE, 2020, pp. 1–6.

[88] A. Roitberg, T. Pollert, M. Haurilet, M. Martin, and R. Stiefelhagen, "Analysis of deep fusion strategies for multi-modal gesture recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[89] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 10, pp. 2684–2701, 2019.

[90] E. Cippitelli, F. Fioranelli, E. Gambi, and S. Spinsante, "Radar and rgb-depth sensors for fall detection: A review," *IEEE Sensors Journal*, vol. 17, no. 12, pp. 3585–3604, 2017.

[91] A. Shahroudy, T.-T. Ng, Y. Gong, and G. Wang, "Deep multimodal feature analysis for action recognition in rgb+ d videos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1045–1058, 2017.

[92] Y. Wang, H. Wu, J. Zhang, Z. Gao, J. Wang, P. S. Yu, and M. Long, "Predrnn: A recurrent neural network for spatiotemporal predictive learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 2208–2225, 2022.

[93] F. Zhu, L. Shao, J. Xie, and Y. Fang, "From handcrafted to learned representations for human action recognition: A survey," *Image and Vision Computing*, vol. 55, pp. 42–52, 2016.

[94] D. Das Dawn and S. H. Shaikh, "A comprehensive survey of human action recognition with spatio-temporal interest point (stip) detector," *The Visual Computer*, vol. 32, pp. 289–306, 2016.

[95] S. S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: a survey," *Artificial intelligence review*, vol. 43, pp. 1–54, 2015.

[96] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," in *Proceedings of the 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*. Bruges, Belgium: i6doc.com, 2013, pp. 437–442.

[97] D. Micucci, M. Mobilio, and P. Napoletano, "Unimib shar: A dataset for human activity recognition using acceleration data from smartphones," *Applied Sciences*, vol. 7, no. 10, p. 1101, 2017.

[98] K. Rezaee, S. M. Rezakhani, M. R. Khosravi, and M. K. Moghimi, "A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance," *Personal and Ubiquitous Computing*, vol. 28, no. 1, pp. 135–151, 2024.

[99] B. Cunling, L. Weigang, and F. Wei, "Skeleton-based human action recognition: History, status and prospects." *Journal of Computer Engineering & Applications*, vol. 60, no. 20, 2024.

[100] V. Mazzia, S. Angarano, F. Salvetti, F. Angelini, and M. Chiaberge, "Action transformer: A self-attention model for short-time pose-based human action recognition," *Pattern Recognition*, vol. 124, p. 108487, 2022.

[101] A. Sepas-Moghaddam and A. Etemad, "Deep gait recognition: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 264–284, 2022.

[102] J. Qi, L. Ma, Z. Cui, and Y. Yu, "Computer vision-based hand gesture recognition for human-robot interaction: a review," *Complex & Intelligent Systems*, vol. 10, no. 1, pp. 1581–1606, 2024.

[103] M. Soori, B. Arezoo, and R. Dastres, "Artificial intelligence, machine learning and deep learning in advanced robotics, a review," *Cognitive Robotics*, vol. 3, pp. 54–70, 2023.

[104] L. Arrotta, C. Bettini, and G. Civitarese, "The marble dataset: Multi-inhabitant activities of daily living combining wearable and environmental sensors data," in *Proceedings of the 18th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous)*. ACM, 2021, pp. 1–10.

[105] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," *21th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN*, vol. 3, pp. 437–442, 2013.

[106] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," *CoRR*, vol. abs/1907.06987, 2019.

[107] G. Vaquette, A. Orcesi, L. Lucat, and C. Achard, "The daily home life activity dataset: a high semantic activity dataset for online recognition," in *2017 12th*

IEEE international conference on automatic face & gesture recognition (FG 2017). IEEE, 2017, pp. 497–504.

[108] activity graz uni, "Volleyball activity dataset dataset," may 2023, visited on 2025-05-25.

[109] M. Fazli, K. Kowsari, E. Gharavi, L. Barnes, and A. Doryab, "Hhar-net: H ierarchical h uman a ctivity r ecognition using neural net works," in *Intelligent Human Computer Interaction: 12th International Conference, IHCI 2020, Daegu, South Korea, November 24–26, 2020, Proceedings, Part I 12*. Springer, 2021, pp. 48–58.

[110] J. Liu, S. Song, C. Liu, Y. Li, and Y. Hu, "A benchmark dataset and comparison study for multi-modal human action analytics," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 2, pp. 1–24, 2020.

[111] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," 2013.

[112] D. Roggen, G. Tröster, P. Lukowicz, A. Ferscha, J. del R. Millán, and R. Chavarriaga, "Opportunity: Towards opportunistic activity and context recognition in sensor networks," *Computer*, vol. 46, no. 2, pp. 36–45, 2013.

[113] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[114] A.-C. Popescu, I. Mocanu, and B. Cramariuc, "Fusion mechanisms for human activity recognition using automated machine learning," in *2020 International Conference on Development and Application Systems (DAS)*. IEEE, 2020, pp. 123–128.

[115] S. Chan, H. Yuan, C. Tong, A. Acquah, A. Schonfeldt, J. Gershuny, and A. Doherty, "Capture-24: A large dataset of wrist-worn activity tracker data collected in the wild for human activity recognition," *arXiv preprint arXiv:2402.19229*, 2024.

[116] N. Sikder and A.-A. Nahid, "Ku-har: An open dataset for heterogeneous human activity recognition," *Pattern Recognition Letters*, vol. 146, pp. 221–228, 2021.

[117] Y. Xu, J. Yang, H. Cao, K. Mao, J. Yin, and S. See, "Arid: A new dataset for recognizing action in the dark," in *Proceedings of the 16th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. IEEE, 2020, pp. 1210–1215.

[118] A. Logacjov, K. Bach, A. Kongsvold, H. B. Bårdstu, and P. J. Mork, "Harth: A human activity recognition dataset for machine learning," *Sensors*, vol. 21, no. 23, p. 7853, 2021.

[119] A. Sucerquia, J. D. López, and J. F. Vargas-Bonilla, "Sisfall: A fall and movement dataset," *Sensors*, vol. 17, no. 1, p. 198, 2017.

[120] F. Niemann, B. Reining, and M. ten Hompel, "Lara: Creating a dataset for human activity recognition in logistics using semantic attributes," *Sensors*, vol. 20, no. 15, p. 4083, 2020.

[121] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, and P. J. M. Havinga, "Complex human activity recognition using smartphone and wrist-worn motion sensors," *Sensors*, vol. 16, no. 4, p. 426, 2016.

[122] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," in *ACM SigKDD Explorations Newsletter*, vol. 12, no. 2. ACM, 2011, pp. 74–82.

[123] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)*, vol. 3. IEEE, 2004, pp. 32–36.

[124] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *CVPR*, 2015.

[125] Z. Wang *et al.*, "The sussex-huawei locomotion dataset for human activity recognition," in *UbiComp*, 2018.

[126] A. Hayat, F. Morgado-Dias, B. Bhuyan, and R. Tomar, "Human activity recognition for elderly people using machine and deep learning approaches," *Information*, vol. 13, no. 6, p. 275, May 2022.

[127] H. Fang, W. Zhou, and H. Li, "End-to-end action quality assessment with action parsing transformer," in *2023 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2023, pp. 1–5.

[128] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, "Deep learning-based human pose estimation: A survey," *ACM Computing Surveys*, vol. 56, no. 1, pp. 1–37, 2023.

[129] W. Wu, Z. Sun, and W. Ouyang, "Revisiting classifier: Transferring vision-language models for video recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 3, 2023, pp. 2847–2855.

[130] D. Kurchaniya and S. Kumar, "D-scan: Dual stream spatiotemporal channel-wise attention network with point-wise convbi-lstm for activity recognition," *IEEE Transactions on Consumer Electronics*, 2024.

[131] W. Li, Y. Wong, A.-A. Liu, Y. Li, Y.-T. Su, and M. Kankanhalli, "Multi-camera action dataset (mcad): a dataset for studying non-overlapped cross-camera action recognition," *CoRR arXiv*, vol. 1607, 2016.

[132] L. Martínez-Villaseñor and H. Ponce, "A concise review on sensor signal acquisition and transformation applied to human activity recognition and human–robot interaction," *International Journal of Distributed Sensor Networks*, vol. 15, no. 6, p. 1550147719853987, 2019.

[133] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, "A review of convolutional neural networks in computer vision," *Artificial Intelligence Review*, vol. 57, no. 4, p. 99, 2024.

[134] F. Setiawan, B. N. Yahya, S.-J. Chun, and S.-L. Lee, "Sequential inter-hop graph convolution neural network (sihgcn) for skeleton-based human action recognition," *Expert Systems with Applications*, vol. 195, p. Article 116566, 2022.

[135] S. A. Khowaja and S. L. Lee, "Skeleton-based human action recognition with sequential convolutional-lstm networks and fusion strategies," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–18, 2022.

[136] L. Wang and D. Suter, "Learning and matching of dynamic shape manifolds for human action recognition," *IEEE Transactions on Image Processing*, vol. 16, no. 6, pp. 1646–1661, 2007.

[137] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, "Skeleton-based human activity recognition using convlstm and guided feature learning," *Soft Computing*, pp. 1–14, 2022.

[138] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[139] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," 2016.

[140] Y. F. Song, Z. Zhang, C. Shan, and L. Wang, "Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, October 2020, pp. 1625–1633.

[141] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Constructing stronger and faster baselines for skeleton-based action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 2, pp. 1474–1488, 2022.

[142] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for rgb-d activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5344–5352.

[143] J. Liu, G. Wang, L. Y. Duan, K. Abdiyeva, and A. C. Kot, "Skeleton-based human action recognition with global context-aware attention lstm networks," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1586–1599, 2018.

[144] S. K. Yadav, A. Luthra, K. Tiwari, H. M. Pandey, and S. A. Akbar, "Arfdnet: An efficient activity recognition & fall detection system using latent feature pooling," *Knowledge-Based Systems*, vol. 239, p. 107948, 2022.

[145] H. Xu, Y. Gao, Z. Hui, J. Li, and X. Gao, "Language knowledge-assisted representation learning for skeleton-based action recognition," *arXiv preprint arXiv:2305.12398*, 2023.

[146] J. Liu, X. Wang, C. Wang, Y. Gao, and M. Liu, "Temporal decoupling graph convolutional network for skeleton-based gesture recognition," *IEEE Transactions on Multimedia*, 2023.

[147] X. Huang, H. Zhou, B. Feng, X. Wang, W. Liu, J. Wang, H. Feng, J. Han, E. Ding, and J. Wang, "Graph contrastive learning for skeleton-based action recognition," *arXiv preprint arXiv:2301.10900*, 2023.

[148] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 20–27.

[149] K. Soomro, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[150] H. Duan, J. Wang, K. Chen, and D. Lin, "Pyskl: Towards good practices for skeleton action recognition," in *Proceedings of the 30th ACM International Conference on Multimedia*. ACM, 2022, pp. 7351–7354.

[151] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: A large video database for human motion recognition," in *2011 International Conference on Computer Vision*, 2011, pp. 2556–2563.

[152] J. Wang and L. Xia, "Abnormal behavior detection in videos using deep learning," *Cluster Computing*, vol. 22, no. Suppl 4, pp. 9229–9239, 2019.

[153] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine vision and applications*, vol. 24, no. 5, pp. 971–981, 2013.

[154] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 1996–2003.

[155] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," *arXiv preprint arXiv:1808.01340*, 2018.

[156] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[157] H. Ullah and A. Munir, "Human activity recognition using cascaded dual attention cnn and bi-directional gru framework," *Journal of Imaging*, vol. 9, no. 7, p. 130, 2023.

[158] A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Action recognition using optimized deep autoencoder and cnn for surveillance data streams of non-stationary environments," *Future Generation Computer Systems*, vol. 96, pp. 386–397, 2019.

[159] A. Hussain, T. Hussain, W. Ullah, and S. W. Baik, "Vision transformer and deep sequence learning for human activity recognition in surveillance videos," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, p. 3454167, 2022.

[160] K. Seemanthini and S. Manjunath, "Human detection and tracking using hog for action recognition," *Procedia computer science*, vol. 132, pp. 1317–1326, 2018.

[161] Z. Zhang, Z. Lv, C. Gan, and Q. Zhu, "Human action recognition using convolutional lstm and fully-connected lstm with different attentions," *Neurocomputing*, vol. 410, pp. 304–316, 2020.

[162] J. Ye, L. Wang, G. Li, D. Chen, S. Zhe, X. Chu, and Z. Xu, "Learning compact recurrent neural networks with block-term tensor decomposition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9378–9387.

[163] C. Dai, X. Liu, and J. Lai, "Human action recognition using two-stream attention based lstm networks," *Applied soft computing*, vol. 86, p. 105820, 2020.

[164] K. Muhammad, Mustaqeem, A. Ullah, A. S. Imran, M. Sajjad, M. S. Kiran, G. Sannino, and V. H. C. de Albuquerque, "Human action recognition using attention based lstm network with dilated cnn features," *Future Generation Computer Systems*, vol. 125, pp. 820–830, 2021.

[165] A. Ullah, K. Muhammad, J. Del Ser, S. W. Baik, and V. H. C. de Albuquerque, "Activity recognition using temporal optical flow convolutional features and multilayer lstm," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 12, pp. 9692–9702, 2018.

[166] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 11, pp. 2740–2755, 2018.

[167] S. Yu, L. Xie, L. Liu, and D. Xia, "Learning long-term temporal features with deep neural networks for human action recognition," *IEEE Access*, vol. 8, pp. 1840–1850, 2019.

[168] C.-Y. Ma, M.-H. Chen, Z. Kira, and G. AlRegib, "Ts-lstm and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition," *Signal Processing: Image Communication*, vol. 71, pp. 76–87, 2019.

[169] M. Majd and R. Safabakhsh, "Correlational convolutional lstm for human action recognition," *Neurocomputing*, vol. 396, pp. 224–229, 2020.

[170] J.-Y. He, X. Wu, Z.-Q. Cheng, Z. Yuan, and Y.-G. Jiang, "Db-lstm: Densely-connected bi-directional lstm for human action recognition," *Neurocomputing*, vol. 444, pp. 319–331, 2021.

[171] J. Xiao, L. Jing, L. Zhang, J. He, Q. She, Z. Zhou, A. Yuille, and Y. Li, "Learning from temporal gradient for semi-supervised action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3252–3262.

[172] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional lstm with cnn features," *IEEE access*, vol. 6, pp. 1155–1166, 2017.

[173] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4768–4777.

[174] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1510–1517, 2017.

[175] X. Long, C. Gan, G. De Melo, J. Wu, X. Liu, and S. Wen, "Attention clusters: Purely attention based local feature integration for video classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7834–7843.

[176] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, and C. G. Snoek, "Videolstm convolves, attends and flows for action recognition," *Computer Vision and Image Understanding*, vol. 166, pp. 41–50, 2018.

[177] Y. Han, P. Zhang, T. Zhuo, W. Huang, and Y. Zhang, "Going deeper with two-stream convnets for action recognition in video surveillance," *Pattern Recognition Letters*, vol. 107, pp. 83–90, 2018.

[178] Y. Zhou, X. Sun, Z.-J. Zha, and W. Zeng, "Mict: Mixed 3d/2d convolutional tube for human action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 449–458.

[179] H. P. Nguyen and B. Ribeiro, "Video action recognition collaborative learning with dynamics via pso-convnet transformer," *Scientific Reports*, vol. 13, no. 1, p. 14624, 2023.

[180] D. He, Z. Zhou, C. Gan, F. Li, X. Liu, Y. Li, L. Wang, and S. Wen, "Stnet: Local and global spatial-temporal modeling for action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 8401–8408.

[181] J. Hsiao, J. Chen, and C. Ho, "Gcf-net: Gated clip fusion network for video action recognition," in *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 2020, pp. 699–713.

[182] Z. Zheng, G. An, D. Wu, and Q. Ruan, "Global and local knowledge-aware attention network for action recognition," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 1, pp. 334–347, 2020.

[183] G. Yang, Y. Yang, Z. Lu, J. Yang, D. Liu, C. Zhou, and Z. Fan, "Sta-tsn: Spatial-temporal attention temporal segment network for action recognition in video," *PloS one*, vol. 17, no. 3, p. e0265115, 2022.

[184] S. Vishwakarma and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," *The Visual Computer*, vol. 29, pp. 983–1009, 2013.

[185] K. Peng, A. Roitberg, K. Yang, J. Zhang, and R. Stiefelhagen, "Delving deep into one-shot skeleton-based action recognition with diverse occlusions," *IEEE Transactions on Multimedia*, vol. 25, pp. 1489–1504, 2023.

[186] M. Ghalan and R. K. Aggarwal, "Novel human activity recognition by graph engineered ensemble deep learning model," *IFAC Journal of Systems and Control*, vol. 27, p. 100253, 2024.

[187] H. Meng, Y. Zhao, Y. Guo, and P. Lv, "Tta-gcn: Temporal topology aggregation for skeleton-based action recognition," in *International Conference on Image and Graphics*. Springer, 2023, pp. 225–237.

[188] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[189] Y.-F. Song, Z. Zhang, C. Shan, and L. Wang, "Richly activated graph convolutional network for robust skeleton-based action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, pp. 1915–1925, 2020.

[190] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[191] W. Shi, D. Li, Y. Wen, and W. Yang, "Occlusion-aware graph neural networks for skeleton action recognition," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 10, pp. 10 288–10 298, 2023.

[192] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[193] F. Heidarivincheh, M. Mirmehdi, and D. Damen, "Beyond action recognition: Action completion in rgb-d data," in *27th British Machine Vision Conference*, 2016.

[194] Z. Yang, K. Li, H. Gan, Z. Huang, and M. Shi, "Hd-gcn: A hybrid diffusion graph convolutional network," *arXiv preprint arXiv:2303.17966*, 2023.

[195] G. Sun, H. Cholakkal, S. Khan, F. Khan, and L. Shao, "Fine-grained recognition: Accounting for subtle differences between similar classes," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 047–12 054.

[196] Z. Chen, H. Wang, and J. Gui, "Occluded skeleton-based human action recognition with dual inhibition training," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 2625–2634.

[197] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 2049–2062, 2018.

[198] W. Cao, R. Wang, M. Fan *et al.*, "Froth image clustering with feature semi-supervision through selection and label information," *International Journal of Machine Learning and Cybernetics*, vol. 12, pp. 2499–2516, 2021.

[199] W. Yang, S. Wang, J. Wu *et al.*, "A low-light image enhancement method for personnel safety monitoring in underground coal mines," *Complex & Intelligent Systems*, vol. 10, pp. 4019–4032, 2024.

[200] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," 2018.

[201] Y. Xu, J. Yang, H. Cao, K. Mao, J. Yin, and S. See, "Arid: A new dataset for recognizing action in the dark," 2022.

[202] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "Enlightengan: Deep light enhancement without paired supervision," *IEEE Transactions on Image Processing*, vol. 30, pp. 2340–2349, 2021.

[203] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1777–1786.

[204] J. López Lobo, J. Del Ser, E. Villar-Rodriguez, N. Bilbao, and S. Salcedo-Sanz, "On the creation of diverse ensembles for nonstationary environments using bio-inspired heuristics," vol. 514, 02 2017, pp. 67–77.

[205] B. Krawczyk, "Active and adaptive ensemble learning for online activity recognition from data streams," *Knowledge-Based Systems*, vol. 138, pp. 69–78, 2017.

[206] K. Abdullah, I. Jegham, M. Mahjoub, and A. Ben Khalifa, "Driver action recognition in low-light conditions: A multi-view fusion framework," 07 2024, pp. 171–176.

[207] Y. Wang and G. Mori, "Hidden part models for human action recognition: Probabilistic versus max margin," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 7, pp. 1310–1323, 2011.

[208] Y. Liu, L. Nie, L. Han, L. Zhang, and D. S. Rosenblum, "Action2activity: recognizing complex activities from sensor data," in *Proceedings of the 24th International Conference on Artificial Intelligence*, ser. IJCAI'15. AAAI Press, 2015, p. 1617–1623.

[209] X. Chang, Y.-L. Yu, Y. Yang, and E. P. Xing, "Semantic pooling for complex event analysis in untrimmed videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 8, pp. 1617–1632, 2017.

[210] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017.

[211] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, " MobileNetV2: Inverted Residuals and Linear Bottlenecks ," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2018, pp. 4510–4520.

[212] K. Sreelakshmi, P. C. Rafeeque, S. Sreetha, and E. S. Gayathri, "Deep bidirectional lstm network for query intent detection," *Procedia Computer Science*, vol. 143, pp. 939–946, 2018, 8th International Conference on Advances in Computing & Communications (ICACC-2018).

[213] A. Radman and S. A. Suandi, "Bilstm regression model for face sketch synthesis using sequential patterns," *Neural Comput. Appl.*, vol. 33, no. 19, p. 12689–12702, Oct. 2021.

[214] Y. Tatsunami and M. Taki, "Sequencer: deep lstm for image classification," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.

[215] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *2013 IEEE International Conference on Computer Vision*, 2013, pp. 3192–3199.

[216] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

[217] K. Soomro and A. Zamir, "Action recognition in realistic sports videos," *Advances in Computer Vision and Pattern Recognition*, vol. 71, pp. 181–208, 01 2014.

[218] F. Yang, Y. Wu, S. Sakti, and S. Nakamura, "Make skeleton-based action recognition model smaller, faster and better," in *Proceedings of the 1st ACM Interna-*

tional Conference on Multimedia in Asia, ser. MMAsia '19. New York, NY, USA: Association for Computing Machinery, 2020.

[219] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, " Two Stream LSTM: A Deep Fusion Framework for Human Action Recognition ," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Los Alamitos, CA, USA: IEEE Computer Society, Mar. 2017, pp. 177–186.

[220] E. P. Ijjina, "Action recognition in sports videos using stacked auto encoder and hog3d features," in *Proceedings of the Third International Conference on Computational Intelligence and Informatics: ICCII 2018*. Springer, 2020, pp. 849–856.

[221] E. P. Ijjina and C. Krishna Mohan, "Hybrid deep neural network model for human action recognition," *Applied Soft Computing*, vol. 46, pp. 936–952, 2016.

[222] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[223] Z. Li, S. Deldari, L. Chen, H. Xue, and F. D. Salim, "Sensorllm: Aligning large language models with motion sensors for human activity recognition," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 354–379.

[224] S. Wang, D. Kim, A. Taalimi, C. Sun, and W. Kuo, "Learning visual grounding from generative vision and language model," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 8057–8067.

[225] S. Ji, X. Zheng, and C. Wu, "Hargpt: Are llms zero-shot human activity recognizers?" in *2024 IEEE International Workshop on Foundation Models for Cyber-Physical Systems & Internet of Things (FMSys)*. IEEE, 2024, pp. 38–43.

[226] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[227] D. R. Beddiar and B. Nini, "Vision based abnormal human activities recognition: An overview," in *2017 8th International Conference on Information Technology (ICIT)*. IEEE, 2017, pp. 548–553.

[228] M. Perez, J. Liu, and A. C. Kot, "Skeleton-based relational reasoning for group activity analysis," *Pattern Recognition*, vol. 122, p. 108360, 2022.

[229] J. C. Nunez, R. Cabido, J. J. Pantrigo, A. S. Montemayor, and J. F. Velez, "Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition," *Pattern Recognition*, vol. 76, pp. 80–94, 2018.

[230] A. Muhamada and A. Mohammed, "Review on recent computer vision methods for human action recognition," *Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 10, no. 4, pp. 361–379, 2021.

[231] C. Chen, R. Jafari, and N. Kehtarnavaz, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 168–172.

[232] J. L. Reyes-Ortiz, D. Anguita, L. Oneto, and X. Parra, "Smartphone-based recognition of human activities and postural transitions," 2015, dataset Version 2.0.

[233] U. A. Akansha, M. Shailendra, and N. Singh, "Analytical review on video-based human activity recognition," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 2016, pp. 3839–3844.

[234] B. Ren, M. Liu, R. Ding, and H. Liu, "A survey on 3d skeleton-based action recognition using learning method," *Cyborg and Bionic Systems*, vol. 5, p. 0100, 2024.

[235] H. Xu, Y. Gao, Z. Hui, J. Li, and X. Gao, "Language knowledge-assisted representation learning for skeleton-based action recognition," *IEEE Transactions on Multimedia*, 2025.

[236] D. Vrontis, M. Christofi, V. Pereira, S. Tarba, A. Makrides, and E. Trichina, "Artificial intelligence, robotics, advanced technologies and human resource management: a systematic review," *Artificial intelligence and international HRM*, pp. 172–201, 2023.

[237] O. Elharrouss, N. Almaadeed, and S. Al-Maadeed, "A review of video surveillance systems," *Journal of Visual Communication and Image Representation*, vol. 77, p. 103116, 2021.

[238] A. A. Khan, A. A. Laghari, and S. A. Awan, "Machine learning in computer vision: A review." *EAI Endorsed Transactions on Scalable Information Systems*, vol. 8, no. 32, 2021.

[239] A. Franco, A. Magnani, and D. Maio, "A multimodal approach for human activity recognition based on skeleton and rgb data," *Pattern Recognition Letters*, vol. 131, pp. 293–299, 2020.

[240] A. Bux, P. Angelov, and Z. Habib, "Vision based human activity recognition: a review," in *Advances in Computational Intelligence Systems: Contributions Presented at the 16th UK Workshop on Computational Intelligence, September 7–9, 2016, Lancaster, UK*. Springer, 2017, pp. 341–371.

[241] A. Manaf and S. Singh, "Computer vision-based survey on human activity recognition system, challenges and applications," in *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*. IEEE, 2021, pp. 110–114.

[242] H. A. Ullah, S. Letchmunan, M. S. Zia, U. M. Butt, and F. H. Hassan, "Analysis of deep neural networks for human activity recognition in videos—a systematic literature review," *IEEE access*, vol. 9, pp. 126 366–126 387, 2021.

[243] T. Singh and D. K. Vishwakarma, "Human activity recognition in video benchmarks: A survey," *Advances in Signal Processing and Communication: Select Proceedings of ICSC 2018*, pp. 247–259, 2019.

[244] M. Babiker, O. O. Khalifa, K. K. Htike, A. Hassan, and M. Zaharadeen, "Automated daily human activity recognition for video surveillance using neural network," in *2017 IEEE 4th international conference on smart instrumentation, measurement and application (ICSIMA)*. IEEE, 2017, pp. 1–5.

Appendices

.1 List of Prime Research Papers Selected in this SLR

ID	Title	Year	Ref.
S1	Radar and RGB-depth sensors for fall detection: A review	2017	[90]
S2	RGB-D-based action recognition datasets: A survey	2016	[64]
S3	From handcrafted to learned representations for human action recognition: A survey	2016	[93]
S4	A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector	2016	[94]
S5	Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions	2019	[63]
S6	Vision based hand gesture recognition for human computer interaction: a survey	2015	[95]
S7	Deep multimodal feature analysis for action recognition in rgb+ d videos	2016	[91]
S8	Predrnn: A recurrent neural network for spatiotemporal predictive learning	2022	[92]
S9	The kinetics human action video dataset	2017	[226]
S10	Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding	2019	[89]
S11	Multi-camera action dataset (MCAD): a dataset for studying non-overlapped cross-camera action recognition	2016	[131]
S12	Vision based abnormal human activities recognition: An overview	2017	[227]
S13	Human activity recognition via hybrid deep learning based model	2022	[18]
S14	Skeleton-based relational reasoning for group activity analysis	2022	[228]
S15	Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition	2018	[229]

ID	Title	Year	Ref.
S16	A concise review on sensor signal acquisition and transformation applied to human activity recognition and human-robot interaction	2019	[132]
S17	Human action recognition systems: A review of the trends and state-of-the-art.	2024	[25]
S18	Vision-based action understanding for assistive healthcare: A short review	2019	[26]
S19	A review on computer vision-based methods for human action recognition	2020	[40]
S20	Human activity recognition using tools of convolutional neural networks: A state-of-the-art review, data sets, challenges, and future prospects	2022	[71]
S21	A survey on hand pose estimation with wearable sensors and computer-vision-based methods	2020	[27]
S22	Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling	2014	[29]
S23	A sound-based human activity monitoring framework for home service robots	2021	[28]
S24	Human activity recognition for elderly people using machine and deep learning approaches	2022	[126]
S25	Vision-based human pose estimation via deep learning: A survey	2023	[31]
S26	Human activity recognition in artificial intelligence framework: A narrative review	2022	[35]
S27	Innohar: A deep neural network for complex human activity recognition.	2019	[36]
S28	Deep learning for computer vision: A brief review	2018	[37]
S29	Human action recognition: A taxonomy-based survey, updates, and opportunities	2023	[38]
S30	Deep learning for sensor-based activity recognition: A survey	2019	[39]
S31	Realtime multi-person 2D pose estimation using part affinity fields.	2017	[42]

ID	Title	Year	Ref.
S32	A survey of vision-based human action evaluation methods.	2019	[43]
S33	Vision and inertial sensing fusion for human action recognition: A review.	2021	[86]
S34	Review on recent computer vision methods for human action recognition	2021	[230]
S35	Tube convolutional neural network (T-CNN) for action detection in videos.	2017	[44]
S36	Tele-evalnet: A low-cost, teleconsultation system for home-based rehabilitation of stroke survivors using multiscale CNN-ConvLSTM architecture	2023	[45]
S37	Attention-guided deep learning framework for movement quality assessment	2023	[**?**]
S38	Multi-region two-stream R-CNN for action detection	2016	[46]
S39	Pecop: Parameter efficient continual pretraining for action quality assessment.	2024	[47]
S40	Skatingverse: A large-scale benchmark for comprehensive evaluation on human action understanding	2024	[48]
S41	Two-path target-aware contrastive regression for action quality assessment.	2024	[50]
S42	LightPRA: A lightweight temporal convolutional network for automatic physical rehabilitation exercise assessment	2024	[51]
S43	EGCN++: A new fusion strategy for ensemble learning in skeleton-based rehabilitation exercise assessment	2024	[52]
S44	Opportunity: Towards opportunistic activity and context recognition in sensor networks.	2013	[112]
S45	Unimib SHAR: A dataset for human activity recognition using acceleration data from smartphones.	2017	[97]
S46	The MARBLE dataset: Multi-inhabitant activities of daily living combining wearable and environmental sensors data.	2021	[104]
S47	Sisfall: A fall and movement dataset.	2017	[119]
S48	LARA: Creating a dataset for human activity recognition in logistics using semantic attributes.	2020	[120]

ID	Title	Year	Ref.
S49	A public domain dataset for human activity recognition using smartphones	2013	[96]
S50	Complex human activity recognition using smartphone and wrist-worn motion sensors.	2016	[96]
S51	UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor	2015	[231]
S52	Smartphone-Based Recognition of Human Activities and Postural Transitions	2015	[232]
S53	Analytical review on video-based human activity recognition	2016	[233]
S54	Vision based hand gesture recognition for human computer interaction: a survey	2015	[95]
S55	Efficient activity recognition using lightweight CNN and DS-GRU network for surveillance applications	2021	[59]
S56	Revisiting classifier: Transferring vision-language models for video recognition	2023	[129]
S57	A survey on 3D skeleton-based action recognition using learning method	2024	[234]
S58	Language knowledge-assisted representation learning for skeleton-based action recognition	2025	[235]
S59	Computer vision-based hand gesture recognition for human-robot interaction: a review	2024	[102]
S60	Self-supervised learning: A succinct review	2023	[54]
S61	Measuring self-regulated learning and the role of AI: Five years of research using multimodal multichannel data	2023	[55]
S62	Artificial intelligence, robotics, advanced technologies and human resource management: a systematic review	2023	[236]
S63	Deep learning-based human pose estimation: A survey	2023	[128]
S64	Self-supervised learning for videos: A survey	2023	[56]
S65	Artificial intelligence, machine learning and deep learning in advanced robotics, a review	2023	[103]
S66	Human action recognition and prediction: A survey	2022	[30]

ID	Title	Year	Ref.
S67	A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions	2021	[62]
S68	Action transformer: A self-attention model for short-time pose-based human action recognition	2022	[100]
S69	A review of video surveillance systems	2021	[237]
S70	Deep gait recognition: A survey	2022	[101]
S71	Machine learning in computer vision: A review	2021	[238]
S72	Sensor-based and vision-based human activity recognition: A comprehensive survey	2020	[80]
S73	A multimodal approach for human activity recognition based on skeleton and RGB data	2020	[239]
S74	NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding	2019	[89]
S75	Deep learning for computer vision: A brief review	2018	[37]
S76	A review on human activity recognition using vision-based method	2017	[66]
S77	Vision based human activity recognition: a review	2017	[240]
S78	Computer vision-based survey on human activity recognition system, challenges and applications	2021	[241]
S79	ConvST-LSTM-Net: convolutional spatiotemporal LSTM networks for skeleton-based human action recognition	2023	[61]
S80	Two stream deep neural network based framework to detect abnormal human activities	2023	[60]
S81	Analysis of deep neural networks for human activity recognition in videos—a systematic literature review	2021	[242]
S82	Human activity recognition in video benchmarks: A survey	2019	[243]
S83	Automated daily human activity recognition for video surveillance using neural network	2017	[244]
S84	A review of state-of-the-art techniques for abnormal human activity recognition	2019	[4]

ID	Title	Year	Ref.
S85	Skeleton-Based Human Action Recognition: History, Status and Prospects.	2024	[99]
S86	D-SCAN: Dual Stream Spatiotemporal Channel-Wise Attention Network With Point-Wise ConvBi-LSTM for Activity Recognition	2024	[130]
S67	A review of convolutional neural networks in computer vision	2024	[133]
S88	A survey on deep learning-based real-time crowd anomaly detection for secure distributed video surveillance	2024	[98]

PUBLICATIONS

Published/ Accepted

SCI-INDEXED JOURNALS

1. Roshni Singh, Abhilasha Sharma."ConvST-LSTM-Net: convolutional spatiotemporal LSTM networks for skeleton-based human action recognition" **International Journal of Multimedia Information Retrieval**, 12,34 (2023). https://doi.org/10.1007/s13735-023-00301-9. (*Impact Factor-2.9).*

2. Roshni Singh, and Abhilasha Sharma."STAD-ConvBi-LSTM: Spatio-temporal attention-based deep convolutional Bi-LSTM framework for abnormal activity recognition." **Journal of Visual Communication and Image Representation** (2025): 104465. https://doi.org/10.1016/j.jvcir.2025.104465 (*Impact Factor-3.1).*

3. Roshni Singh, and Abhilasha Sharma. "Occluded skeleton-based multi-stream model using Part-Aware Spatial–Temporal Graph Convolutional Network for human activity recognition." **Engineering Applications of Artificial Intelligence** 156(2025): 111183. https://doi.org/10.1016/j.engappai.2025.111183 (*Impact Factor-8.0).*

Communicated/Under Review

SCI-INDEXED JOURNALS

1. Roshni Singh, Abhilasha Sharma, "Deep Learning for Human Activity Recognition: A Systematic Review of a Decade of Progress", in **Knowledge and Information Systems**. (Under Review) (*Impact Factor-2.5).*

2. Roshni Singh, Abhilasha Sharma, "Spatio-Temporal Siamese Ensemble with Multi-Level Feature Fusion for Video-Based Person Re-Identification", in **Expert Systems with Applications**. (Communicated) (*Impact Factor-7.5)*.

3. Roshni Singh, Abhilasha Sharma, " Entropy-Aware Cricket Activity Recognition Using CNN and DBSCAN Clustering Computational Statistics and Data Analysis", in **Journal of Science in Sport and Exercise**. (Under Revision) (*Impact Factor-1.3)*.

CONFERENCES

SCOPUS-INDEXED JOURNALS

1. Roshni Singh, Abhilasha Sharma, "MV-DBiLSTM: An Enhanced Human Activity Recognition for Smart Surveillance Systems Using a Deep BiLSTM," 2nd International Conference on Recent Advances in Engineering and Computer Applications (ICRAECA-2025). **(Scopus, Accepted and Presented 27 Jan 2025)**.

2. Roshni Singh, and Abhilasha Sharma. "Activity Recognition in Dynamic Environments Using Image Enhancement and Vision Transformers with DETR." International Conference On Innovative Computing And Communication. Singapore: Springer Nature Singapore, 2025. https://doi.org/10.1007/978-981-96-7707-8_17

International Journal of Multimedia Information Retrieval (2023) 12:34
https://doi.org/10.1007/s13735-023-00301-9

REGULAR PAPER

ConvST-LSTM-Net: convolutional spatiotemporal LSTM networks for skeleton-based human action recognition

Abhilasha Sharma[1] · **Roshni Singh**[1]

Received: 11 February 2023 / Revised: 10 August 2023 / Accepted: 24 September 2023 / Published online: 27 October 2023

Abstract
Human action recognition (HAR) emphases on perceiving and identifying the action behavior done by humans within an image/video. The HAR activities include motion patterns and normal or abnormal activities like standing, walking, sitting, running, playing, falling, fighting, etc. Recently, it sparks the attention of researchers especially in 3D skeleton sequence. The actions of human can be represented via sequence of motions of skeletal keyjoints, although not all the skeleton keyjoints are informative in nature. Various approaches for HAR are used like LSTM, ConvLSTM, Conv-GRU, ST-LSTM, etc. Thus far, ST-LSTM approaches have shown tremendous performance in 3D skeleton sequence tasks but the detection of irrelevant keyjoints produce noise that deteriorates the performance of the model. So, the intent is to bring attention toward improving the efficacy of the model by focusing on informative keyjoint coordinates only. Therefore, the research paper introduces a new class of spatiotemporal LSTM approaches named as ConvST-LSTM-Net (convolutional spatiotemporal long short-term memory network) for skeleton-based action recognition. The prime focus of proposed model is to identify the informative keyjoints in each frame. The result of extensive experimental analysis exhibits that ConvST-LSTM-Net outperforms the state-of-the-art models on various benchmarks dataset, viz. NTU RGB + D 60, UT-Kinetics, UP-Fall Detection, UCF101, and HDMB51 for skeleton sequence data.

Keywords Activity recognition · LSTM · Spatiotemporal · Skeleton sequence · ConvLSTM

1 Introduction

Human action recognition has turn out to be a prominent & diligent research area in computer vision and image processing, which includes classification and recognition of normal & abnormal human activities of daily routine. It belongs to the automated recognition of human activity (normal & abnormal) in various application areas via analyzing the sequence of observations. Nowadays, crowded places with normal and abnormal activities are familiar due to population increase that turn toward suspicious activities. So, HAR has become an essential part in the automatic interpretation of human environment interaction in various online-offline applications such as auto-driving, intelligent surveillance [1–5], smart-gadgets analysis [6], object detection & tracking [7], video retrieval [8], and assisted daily living. Other HAR applications firmly coupled with the daily activities such as motion analysis [9–12], pose motion analysis [13, 14], health monitoring [15], classification or detection of actions or motions [16], and understanding human action behavior [17]. By recognizing and analyzing the human actions from the videos, one can clearly distinguish between normal and abnormal behaviors that can make significant improvements in public safety. Withal, HAR remains an ambitious challenge due to its clutter backgrounds, slight interclass segregation, and wide intra-class deviation. The main thing to recognize high accuracy & efficiency is to conquer both static appearances within each frame of the videos as well as temporal relationships throughout the multiple frames generated via videos. Some applications such as monitoring suspicious detection and early reporting for fall detection are also considered in human activity recognition. However, various techniques are there for the representation of human action based on motion, such as RGB-based

✉ Roshni Singh
roshnisingh1815@gmail.com

Abhilasha Sharma
abhi16.sharma@gmail.com

[1] Department of Software Engineering, Delhi Technological University, Shahbad Daulatpur, Delhi 110042, India

J. Vis. Commun. Image R. 110 (2025) 104465

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Full Length Article

STAD-ConvBi-LSTM: Spatio-temporal attention-based deep convolutional Bi-LSTM framework for abnormal activity recognition☆

Roshni Singh, Abhilasha Sharma *

Delhi Technology University, Shahbad, Delhi, 110042, India

ARTICLE INFO

Keywords:
Activity recognition
Video surveillance
Convolutional neural networks
Bi-LSTM
Attention mechanism

ABSTRACT

Human Activity Recognition has become significant research in computer vision. Real-time systems analyze the actions to endlessly monitor and recognize abnormal activities, thereby enlightening public security and surveillance measures in real-world. However, implementing these frameworks is a challenging task due to miscellaneous actions, complex patterns, fluctuating viewpoints or background cluttering. Recognizing abnormality in videos still needs exclusive focus for accurate prediction and computational efficiency. To address these challenges, this work introduced an efficient novel spatial–temporal attention-based deep convolutional bidirectional long short-term memory framework. Also, proposes a dual attentional convolutional neural network that combines CNN model, bidirectional-LSTM and spatial–temporal attention mechanism to extract human-centric prominent features in video-clips. The result of extensive experimental analysis exhibits that STAD-ConvBi-LSTM outperforms the state-of-the-art methods using five challenging datasets, namely UCF50, UCF101, YouTube-Action, HMDB51, Kinetics-600 and on our Synthesized Action dataset achieving notable accuracies of 98.8%, 98.1%, 81.2%, 97.4%, 88.2% and 96.7%, respectively.

1. Introduction

Human activity recognition (HAR) is often associated with the procedure of identifying and recognizing human activities or behavior or actions. It has become a significant research area because of its robust nature towards the active appearances of real-world situations such as diverse illumination conditions, background clutter, variable camera & its viewpoints, and variations in human body scale. An activity may be defined as the motion executed by a human, occurring over a relatively short period and involving multiple body parts [1]. HAR is an area that analyzes the hidden consecutive pattern of human activity and predicts its state of action, whether it is normal or abnormal, based on perceptual situations, as in the video frames. In contrast to images, video contain more information. Still some challenges arise in videos due to camera movements, scaling variations, human posture and fluctuations in illumination conditions, significantly augmenting the complexity of activity recognition in video sequencing data [2–4]. Thus, for the retrieval of activities based on motion, lots of HAR applications are there, which may include video summarization [5], human–computer interaction [6], education [7], healthcare [8], surveillance [9,10] and sports [11]. In video datasets, a grouping of different human parts in motion is termed as human activity, likewise body+hand+ arms+legs+ face or a grouping of all body parts movements. For example, walking encompasses quick movement of body joints with hands and legs; Likewise, picking up an object consists of to and fro movement of arms, etc. Formerly, the research based on abnormal detection of human patterns was entirely attentive to the activities performed by a single or multiple humans, including a single object, but in a controlled setting [12]. Currently, research emphasizes addressing more problems and authentic scenarios, which include occlusion & cluttered backgrounds, inter-intra variations, different viewpoints, etc. Based on the human patterns, the HAR methods are broadly categorized into three graded viz.: (i) handcrafted feature-based HAR [10,13–15], (ii) deep learning-based HAR [16–21], and (iii) attention mechanism-based HAR [22–24] methods.

The handcrafted feature methods are the manual extraction of features by the engineers, explicitly designed for certain scenarios based on their perceptual complexity. Scale-Invariant Feature Transform (SIFT) [25], Histograms of Oriented Gradients (HOG) [26], Histogram of oriented gradients for 3D (HOG3D) [27], Local Binary Pattern (LBP) [28], histogram optical flow (HOF) [29], Global Image Structure (GIST) [30], Gabor Filter (GF) [31], speeded-up robust feature (SURF) [32] descriptors are some examples. However, these methods could be more productive while addressing the long-term temporal feature dependencies for complex or multi-dependent scenario. Deep

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
E-mail addresses: roshnisingh1815@gmail.com (R. Singh), abhilasha_sharma@dce.ac.in (A. Sharma).

https://doi.org/10.1016/j.jvcir.2025.104465
Received 17 September 2024; Received in revised form 5 January 2025; Accepted 14 April 2025
Available online 28 April 2025

Engineering Applications of Artificial Intelligence 156 (2025) 111183

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Occluded skeleton-based multi-stream model using Part-Aware Spatial–Temporal Graph Convolutional Network for human activity recognition

Roshni Singh, Abhilasha Sharma *

Department of Software Engineering, Delhi Technological University, New Delhi, 110042, Delhi, India

ARTICLE INFO

Keywords:
Occlusion
Skeleton-based human activity recognition
Spatial–temporal graph convolutional network
Inhibition strategy
Predication score map

ABSTRACT

Human activity recognition using skeleton data has engrossed significant research attention in pattern recognition due to its broad applications. However, occlusion remains a major challenge in activity recognition. In this paper, we propose a multi-stream part-aware occluded skeleton-based graph convolutional network designed to improve predictions in the presence of occlusions. The model consists of three key modules: the Input Inhibition Module for Skeleton Sequences, which handles incomplete or occluded skeleton data; the Part-Aware Spatial–Temporal Graph Convolutional Network, which captures spatial–temporal dependencies among human body key joints and the Predicted Score Inhibition, which refines the output by mitigating the effects of noisy data. By integrating these components, the model enhances robustness in occluded scenarios. The experiments demonstrate that the proposed method outperforms state-of-the-art models on several benchmark datasets, achieving a 6% improvement in recognition accuracy compared to previous approaches. Additionally, we extracted multi-modal features to construct more discriminative features, such as key-joint coordinates, relative coordinates, and temporal differences.

1. Introduction

Human activity recognition (HAR) is the possession of identifying actions or behaviors of humans grounded on observations. The substantial application of HAR in various areas such as video-image retrieval, robotics, pattern analysis, intelligent surveillance, entertainment, human–computer interaction, and security. Skeleton-based approaches provide a more accurate structure and relevant information than other modalities like RGB, Optical Flow, and deep learning-based Graph Convolutional Networks (GCN). It is highly robust to variations in illumination, intensity, and different backgrounds, combined with its low-dimensional feature representation, significantly conserving computing resources. Due to progress in depth sensor technology, skeleton data have become increasingly accurate and accessible. These improvements have contributed considerably to the growing status of skeleton-based HAR and become a preferred choice in the research areas of computer vision (Vishwakarma and Agrawal, 2013). Deep learning (DL) has attained significant advancements in research, especially in domains like image recognition, voice recognition, gesture recognition, natural language processing, pattern recognition, etc. Alkatheiri (2022) and Nasir et al. (2022). Earlier, DL-based methods such as Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), or Gated Recurrent Unit (GRU) are commonly employed for spatiotemporal contextual feature information within skeleton data because of their ability to steer dynamic dependencies in sequential data effectively (Kong and Fu, 2022; Peng et al., 2020; Zhang et al., 2020; Zam et al., 2024; Khodabandelou et al., 2023). CNNs also exhibit excellent learning abilities by transforming skeleton sequence data into pseudo-images, enabling the network to effectively capture and analyze spatial–temporal features (Ehatisham-Ul-Haq et al., 2019; Zhou et al., 2016). Since skeleton input sequence can be regarded as a graph data type, it offers a more accurate and comprehensive representation, such as graph structure in the form of edges and body key-joints. GCN-based methods have become progressively popular and attained significant achievement (Peng et al., 2023; Ghalan and Aggarwal, 2024; Han et al., 2024; Meng et al., 2023; Kurchaniya and Kumar, 2024; Sharma and Singh, 2023). However, these methods often fail while addressing common problems such as occlusion and multi-interaction. When an object obstructs critical body key-joints, the recognition capabilities of these models are substantially compromised. Despite the rapid advancements in skeleton-based HAR techniques, improving the robustness of models, especially in occluded environments, remains a significant

* Corresponding author.
E-mail address: abhilasha_sharma@dce.ac.in (A. Sharma).

https://doi.org/10.1016/j.engappai.2025.111183
Received 5 November 2024; Received in revised form 15 April 2025; Accepted 13 May 2025
Available online 3 June 2025

IFERP Academy
Explore Your Research Journey...

CERTIFICATE
OF PRESENTATION

LJ University
University with a Difference

2nd International Conference on Recent Advances in Engineering and Computer Applications (ICRAECA-2025)

24th - 25th January 2025 | Gujarat, India

Certificate No:

This is to Certify that **Ms. Roshni Singh**

.....presented their worthy

Paper titled *MV-DBiLSTM: An Enhanced Human Activity Recognition for Smart Surveillance Systems Using a Deep BiLSTM Framework*

during the "2nd International Conference on Recent Advances in Engineering and Computer Applications (ICRAECA-2025)" Organized by LJ School of Computer Applications LJ University in association with IFERP Academy held on 24th - 25th January 2025 at Gujarat, India.

Mr. Alok Manke
Director
LJ School of Computer Applications
LJ University

Mr. Siddth Kumar Chhajer
MD & Founder, IFERP
Technoarete Group

Mr. Rudra Bhanu Satpathy
CEO & Founder, IFERP
Technoarete Group

Academic Partner
SHINAWATRA UNIVERSITY
FOSTERING INNOVATION

ICICC-2025

INTERNATIONAL CONFERENCE ON INNOVATIVE COMPUTING AND COMMUNICATION (ICICC-2025)

Certificate

This is to certify that **Prof. / Dr. / Mr. / Ms. Roshni singh** is a presenter/co-author of the paper titled **Activity Recognition in Dynamic Environments Using Image Enhancement and Vision Transformers with DETR** in the **8th International Conference on Innovative Computing and Communication (ICICC-2025)**, organized by Shaheed Sukhdev College of Business Studies, University of Delhi, New Delhi, India in association with the National Institute of Technology Patna, India and the University of Valladolid, Spain on **14th-15th February 2025**.

Poonam Verma	**Prabhat Kumar**	**A K Singh**
Principal, SSCBS, University of Delhi, New Delhi	General Chair National Institute of Technology, Patna	Technical Program Chair National Institute of Technology, Kurukshetra

4% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Cited Text
- Small Matches (less than 10 words)

Exclusions

- 3 Excluded Sources

Match Groups

169 Not Cited or Quoted 4%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

1 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

2% Internet sources
3% Publications
2% Submitted works (Student Papers)

Integrity Flags

1 Integrity Flag for Review

Replaced Characters
125 suspect characters on 31 pages
Letters are swapped with similar characters from another alphabet.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Author Biography

Roshni Singh
Research Scholar,
Department of Software & Engineering
Delhi Technological University, Delhi, India
Email: roshnisingh1815@gmail.com

Roshni Singh received her B.Tech. degree in Computer Science & Engineering in 2013 from Punjab Technical University, Punjab. She has completed her Master of Technology from Madan Mohan Malaviya University of Technology, Gorakhpur, U.P in 2017. She is pursuing her PhD from Delhi Technological University, New Delhi, India. She has 3 years of teaching experience and 2 years of Industrial Experience in IBM Cloud. Her research interests include computer vision, image processing, pattern recognition, and artificial intelligence.