UNOBTRUSIVE CONTEXT-AWARE HUMAN IDENTIFICATION AND ACTION RECOGNITION SYSTEM FOR SMART ENVIRONMENTS

Thesis
Submitted in partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

by

RASHMI M

DEPARTMENT OF INFORMATION TECHNOLOGY

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA
SURATHKAL, MANGALORE - 575 025, INDIA

AUGUST, 2023

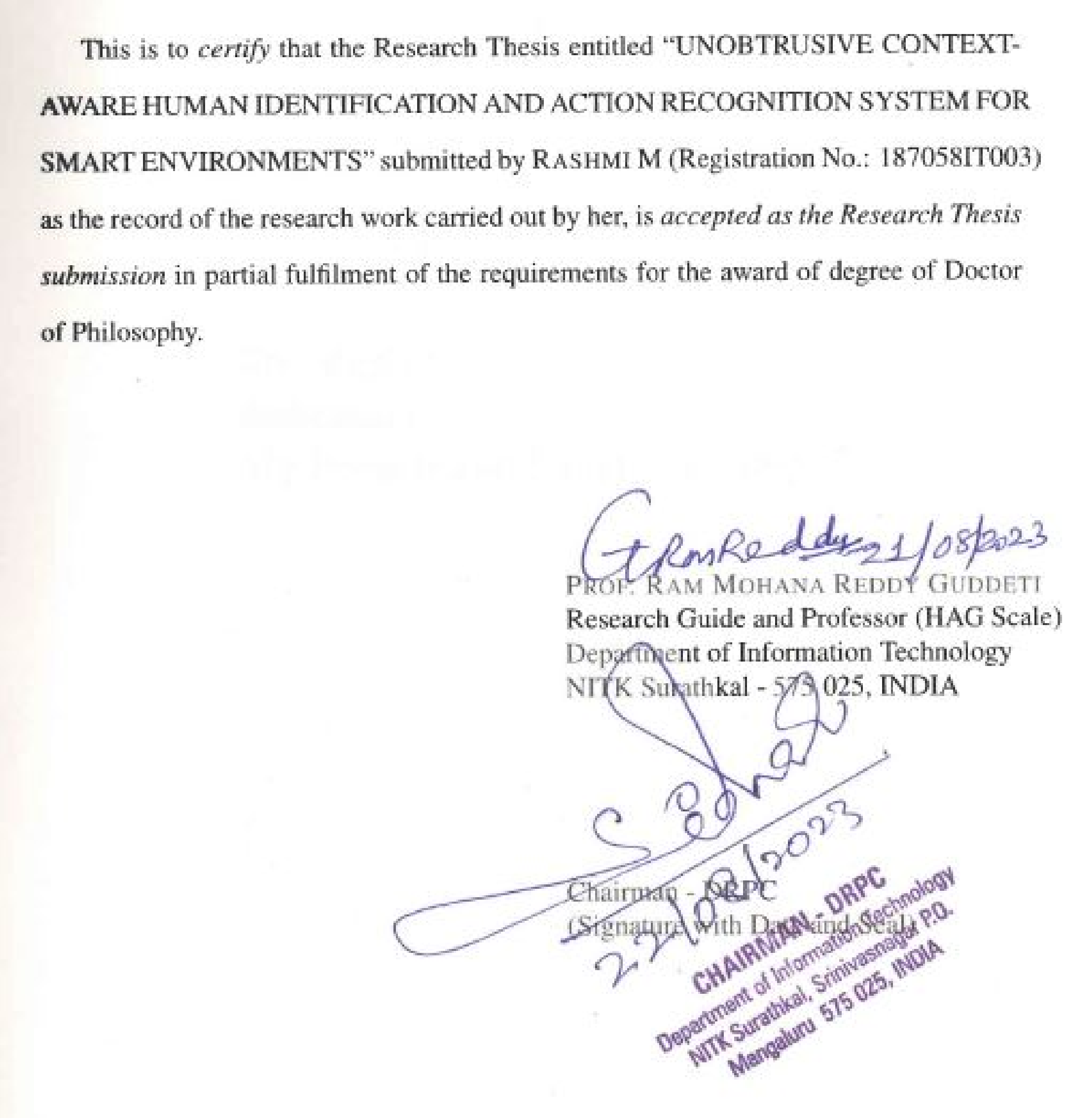
Declaration

I hereby *declare* that the Research Thesis entitled "UNOBTRUSIVE CONTEXT-AWARE HUMAN IDENTIFICATION AND ACTION RECOGNITION SYSTEM FOR SMART ENVIRONMENTS" which is being submitted to the National Institute of Technology Karnataka, Surathkal in partial fulfilment of the requirements for the award of the Degree of Doctor of Philosophy in Information Technology is a *bonafide report of the research work carried out by me*. The material contained in this thesis has not been submitted to any University or Institution for the award of any degree.

Place : NITK Surathkal
Date : 21/08/2023

RASHMI M
Registration No.: 187058IT003
Department of Information Technology

Certificate



This is to *certify* that the Research Thesis entitled "UNOBTRUSIVE CONTEXT-AWARE HUMAN IDENTIFICATION AND ACTION RECOGNITION SYSTEM FOR SMART ENVIRONMENTS" submitted by RASHMI M (Registration No.: 187058IT003) as the record of the research work carried out by her, is *accepted as the Research Thesis submission* in partial fulfilment of the requirements for the award of degree of Doctor of Philosophy.

GRmReddy 21/08/2023
PROF. RAM MOHANA REDDY GUDDETI
Research Guide and Professor (HAG Scale)
Department of Information Technology
NITK Surathkal - 575 025, INDIA

22/08/2023
Chairman - DRPC
(Signature with Date and Seal)

CHAIRMAN - DRPC
Department of Information Technology
NITK Surathkal, Srinivasnagar P.O.
Mangaluru 575 025, INDIA

This thesis is
dedicated to
My Parents and Family Members

Acknowledgements

First and foremost, I express my profound gratitude and appreciation to my mentor and research guide, Prof. Ram Mohana Reddy Guddeti, Professor (HAG Scale), Department of Information Technology, NITK Surathkal, for his guidance, insightful feedback, inspiration and unwavering commitment to helping me succeed in my research work. His prompt advice and decisive actions led to completion of this research work. I consider myself extremely fortunate to have him as my supervisor.

My sincere appreciation to Dr. Shashidhar G. Koolagudi, Dept. of Computer Science & Engineering and Dr. Jaidhar C. D., Dept. of Information Technology, who have served as members of RPAC, for their insightful comments and recommendations. I am grateful to Dr. Biju R. Mohan for his support during my doctoral studies. I would also like to thank all of the teaching, technical, administrative, and non-teaching staffs who have been extremely kind and helpful throughout my research work.

I am indebted to my deceased grandparents, father, Late Govinda Naik, and father-in-law, Late Krishna Naik, for their blessings and inspiration to pursue a doctorate. I owe a great debt of gratitude to my dear mother, Mrs. Susheela M, and mother-in-law Mrs. Gopi, for their encouragement and support throughout my research studies, which allowed me to balance family responsibilities and research work. Also, I am grateful to my brother and sister for their encouragement throughout my studies.

I am privileged to thank my husband, Dr. Dinesh Naik, who has tremendously supported my research efforts. Also, thanks to my daughter Twisha D N and son Thejas D N for their unwavering love and support. Nothing would have been accomplished without their support and sacrifices.

I am incredibly grateful to Dr. Ashwin T. S. and Dr. Natesha B. V. for their encouragement, insightful suggestions, and help at various stages of my research journey. In addition, I would like to thank all of my co-researchers and friends who have been helpful and supportive throughout the ups and downs of this research work and made this experience enjoyable and memorable.

Finally, I express my deepest gratitude to NITK Surathkal, Karnataka, India, the most exciting location, for making this research more memorable and possible by providing cutting-edge facilities to conduct research work.

Rashmi M

Abstract

A smart environment has the ability to securely integrate multiple technological solutions to manage its assets, such as the information systems of local government departments, schools, transportation networks, hospitals, and other community services. They utilize low-power sensors, cameras, and software with Artificial Intelligence to continuously monitor the system's operation. Smart environments require appropriate monitoring technologies for a secure living environment and efficient management.

Global security threats have produced a considerable demand for intelligent surveillance systems in smart environments. Consequently, the number of cameras deployed in smart environments to record the happenings in the vicinity is increasing rapidly. In recent years, the proliferation of cameras such as Closed Circuit Television (CCTV), depth sensors, and mobile phones used to monitor human activities has led to an explosion of visual data. It requires considerable effort to interpret and store all of this visual data. Numerous applications of intelligent environments rely on the content of captured videos, including smart video surveillance to monitor human activities, crime detection, intelligent traffic management, human identification, etc.

Intelligent surveillance systems must perform unobtrusive human identification and human action recognition to ensure a secure and pleasant life in a smart environment. This research thesis presents various approaches using advanced deep learning technology for unobtrusive human identification and human action recognition based on visual data in various data modalities. This research thesis explores the unobtrusive identification of humans based on skeleton and depth data. Also, several methods for recognizing human actions using RGB, depth, and skeleton data are presented.

Initially, a domain-specific human action recognition system employing RGB data for a computer laboratory in a college environment is introduced. A dataset of human actions particular to the computer laboratory environment is generated using spontaneous video data captured by cameras installed in laboratories. The dataset contains several instances of five distinct human actions in college computer laboratories. Also, human action recognition system based on transfer learning is presented for locating and recognizing multiple human actions in an RGB image.

Human action recognition systems based on skeleton data is developed and evaluated on publicly available datasets using benchmark evaluation protocols and metrics. The skeleton data-based action recognition mainly concentrates on the 3D coordinates of various skeleton joints of the human body. This research thesis presents several efficient action representation methods from the data sequence in skeleton frames. A skeleton data-based human action recognition system places the skeleton joints in a specific order, and the distance between joints is extracted as features. A multi-layer deep learning model is proposed to learn the features and recognize human actions.

Human gait is one of the most useful biometric features for human identification. The vision-based gait data allows human identification unobtrusively. This research thesis presents deep learning-based human identification systems using gait data in skeleton format. We present an efficient feature extraction method that captures human skeleton joints' spatial and temporal features during walking. This specifically focuses on the features of different gait events in the entire gait cycle. Also, deep learning models are developed to learn these features for accurate human identification systems. The developed models are evaluated on publicly available single and multi-view gait datasets using various evaluation protocols and performance metrics.

In addition, multi-modal human action recognition and human identification systems are developed using skeleton and depth data. This presents efficient image representations of human actions from the sequence of frames in skeleton and depth data formats. Various deep learning models using CNN, LSTM, and advanced techniques such as Attention is presented to extract and learn the features from image representation of the actions. Also, another work presents a method focusing on overlapping sub-actions of action in depth and skeleton format for action representation and feature extraction. In addition, the image representation of the gait cycle in skeleton and depth data, along with a deep learning model, is proposed. Multi-stream deep learning models are proposed to learn features from multi-modal data for human action recognition and human identification. In addition, various score fusion operations are proposed to merge the results from multiple streams of deep learning models to ensure efficient performance. The developed systems are evaluated on publicly available multi-modal datasets for human actions and human gait using standard evaluation protocols.

Keywords: Attention, CNN, Deep learning, Depth data, Human action recognition, Human identification, LSTM, Multi-modal, Score fusion, Skeleton data, Smart environments, Smart surveillance.

Contents

List of Abbreviations **vii**

List of Figures **xi**

List of Tables **xv**

1 Introduction **1**

1.1 Overview of Human Activity/Action Recognition 2

1.1.1 Categories of Human Actions 2

1.1.2 Categories of HAR Systems 3

1.2 Overview of Gait-based Human Identification 7

1.2.1 Gait Cycle Components 8

1.3 Approaches for Gait Recognition 9

1.4 Motivation 10

1.5 Organization of Thesis 11

1.6 Summary 14

2 Literature Review **15**

2.1 Human Action Recognition 15

2.1.1 RGB Image-based HAR 16

2.1.2 HAR using Skeleton Data 19

2.1.3 HAR using Depth Data 23

2.1.4 Multi-modal HAR 25

2.1.5 Datasets for HAR 27

2.2 Human Identification 29

2.2.1 2D Gait-based Human Identification 31

2.2.2 3D Gait-based Human Identification 31

2.2.3 3D Gait Datasets for Human Identification 34

2.3 Outcome of Literature Review 35

2.4 Problem Statement 37

2.5 Research Objectives 37

2.6 Summary 38

3 Context-aware Human Action Recognition System for Computer Laboratories of Smart Campus 39

3.1 Smart Campus 39

3.2 Proposed Methodology 40

3.2.1 Dataset Preparation 42

3.2.1.1 Collect Images with Human Actions 42

3.2.1.2 Image Pre-processing 42

3.2.1.3 Image Augmentation 42

3.2.1.4 Data Annotation 43

3.2.2 Training Deep Learning Model 43

3.2.2.1 Overview of YOLOv3 43

3.2.3 Action Recognition 44

3.3 Details about Dataset 46

3.4 Experiments, Results, and Analysis 47

3.4.1 Testing 49

3.4.1.1 Intersection over Union 50

3.4.1.2 Kappa Coefficient 51

3.4.1.3 Analysis of Detected Actions 53

3.4.2 Limitations of the Work 57

3.5 Summary 58

4 Single-view Human Action Recognition System using Skeleton Data 59

4.1 Proposed Methodology 60

4.1.1 Feature Extraction and Feature Fusion 60

4.1.2 Neural Network Model 63

4.1.3 Score Fusion 63

4.2 Experiments, Results, and Analysis 64

4.2.1 Dataset 64

4.2.2 Experiments and Results 65

4.2.3 Comparison with Existing Works 70

4.2.4 Computational Cost Analysis 70

4.3 Limitations 71

4.4 Summary 71

5 Human Identification System using Single-/ Multi-view Skeleton-based Gait Data 73

5.1 LSTM Model-based HI System using 3D Skeleton-based Gait Data 75

5.1.1 Proposed Methodology 75

5.1.1.1 Pre-processing 76

5.1.1.2 Gait Cycle Extraction 78

5.1.1.3 Gait Event Frame Grouping 80

5.1.1.4 Feature Extraction 80

5.1.1.5 LSTM Based Deep Learning Model 83

5.1.1.6 Long Short-Term Memory and Gated Recurrent Unit 84

5.1.1.7 Score Fusion 86

5.1.2 Experiments, Results, and Analysis 88

5.1.2.1 Datasets 88

5.1.2.2 Evaluation Protocols 89

5.1.2.3 Evaluation Metrics 90

5.1.2.4 Training Parameters 91

5.1.2.5 Experiments on Single-view Dataset 91

5.1.2.6 Experiments on Multi-view Dataset 99

5.1.2.7 Ablation Experiments 101

5.1.2.8 Statistical Analysis 102

5.1.2.9 Computational Cost Analysis and Discussion . . . 105

5.2 Attention-guided Residual Deep Learning Model and Gait Event-specific Features for Human Identification from Skeleton-based Gait Data . 107

5.2.1 Proposed Methodology 107

5.2.1.1 Pre-processing 109

5.2.1.2 Feature Extraction 109

5.2.1.3 Person Identification 109

5.2.2 Experiments, Results, and Analysis 113

5.2.2.1 Dataset 113

5.2.2.2 Evaluation Protocols and Metrics 113

5.2.2.3 Training Parameters 114

5.2.2.4 Experiments Considering Single-view Scenario . 114

5.2.2.5 Experiments Considering Multi-view Scenario . . 117

5.2.2.6 Ablation Experiments 118

5.2.2.7 Cumulative Match Characteristic Curve 121

5.2.2.8 Further Statistical Analysis 122

5.2.2.9 Computational Complexity Analysis 125

5.3 Limitations 125

5.4 Summary 126

6 Single-/ Multi-view Human Action Recognition and Human Identification using Fusion of Data Modalities 127

6.1 Skeleton and Depth Data-driven Multi-stream DL Model for Single-view HAR 128

6.1.1 Proposed Methodology 128

6.1.1.1 Pre-processing 129

6.1.1.2 Attention-guided Multi-stream CNN + LSTM Model 131

6.1.1.3 Score Fusion 134

6.1.2 Experiments, Results, and Analysis 134

6.1.2.1 Datasets 134

6.1.2.2 Evaluation Protocol and Metrics 135

6.1.2.3 Experiments and Results 135

6.1.2.4 Comparison with Existing Works 136

6.2 Multi-stream Attention-guided Deep Networks with Skeleton and Depth Data from Overlapping Sub-actions for Single-/Multi-view HAR . . 139

6.2.1 Proposed Methodology 140

6.2.1.1 Pre-processing 141

6.2.1.2 Multi-stream Attention-based Deep Neural Network Model 143
6.2.1.3 Score Fusion 149
6.2.2 Experiments, Results, and Analysis 151
6.2.2.1 Dataset 151
6.2.2.2 Evaluation Protocols and Metrics 151
6.2.2.3 Experiments and Results 152
6.2.2.4 Comparison with State-of-the-art Works 159
6.3 HI System using Fusion of Gait Data in Skeleton and Depth Modality 161
6.3.1 Proposed Methodology 161
6.3.1.1 Pre-processing 162
6.3.1.2 Deep Learning Model 164
6.3.2 Experiments, Results, and Analysis 165
6.3.2.1 Dataset 165
6.3.2.2 Experiments, and Results 165
6.4 Limitations 165
6.5 Summary 166

7 Conclusions and Future Directions **167**
7.1 Conclusions 167
7.2 Future Directions 170

References **173**

List of Abbreviations

Abbreviation	Meaning	Page No.
2D	Two Dimension	10
3D	Three Dimension	4
ANN	Artificial Neural Network	18
ANOVA	Analysis of Variance	102
AP	Average Precision	54
BN	Batch Normalization	83
BpLF	Body-part Length Features	92
CCTV	Closed Circuit Television	40
CMC	Cumulative Match Characteristic	73
CNN	Convolutional Neural Network	7
$C-S$	Cross-Subject	135
CSGEI	Color-coded Skeleton Gait Energy Image	163
CV	Computer Vision	1
$C-V$	Cross-View	151
DB-LSTM	Bi-directional LSTM	18
DFS	Depth First Search	13
DistJ	Distance covered by Joint	109
DL	Deep Learning	4
DME	Depth Motion Energy	143
DMM	Depth Motion Maps	24
DSTI	Depth Spatio-Temporal Image	131
EDMM	Enhanced Depth Motion Map	143
FP	False Positive	53
GCN	Graph Convolutional Network	20
GEI	Gait Energy Image	31
GRU	Gated Recurrent Units	7
HAR	Human Activity/Action Recognition	2

HBMI	History of Binary Motion Image	18
HI	Human Identification	2
HOG	Histograms of Oriented Gradients	24
HOI	Human-Object Interaction	17
IAS-Lab	IAS-Lab RGBD-ID	113
IC	Initial Contact	8
InJAF	Inter-Frame Joint Angle Features	83
InJDF	Inter-Frame Joint Distance Features	81
IoT	Internet of Things	39
IoU	Intersection over Union	50
ISW	Initial Swing	9
KGBD	Kinect Gait Biometry Dataset	88
KNN	K-Nearest Neighbors	32
KS20	VisLab Multi-View Kinect Skeleton	88
LGSD	Local Graphical Skeleton Descriptor	34
LR	Loading Response	9
LSTM	Long Short-Term Memory	7
mAP	mean Average Precision	54
MFSS	Multilevel Frame Select Sampling	25
MJDLR	Mean of Distance between Joints in Left and Right body part	109
ML	Machine Learning	33
MLB	Mean Length of Bones	109
MS-EDMM	Multiple Sub-action Enhanced Depth Motion Map	143
MST	Mid Stance	9
MSW	Mid Swing	9
MTCNN	Multitask Convolutional Neural Network	21
NN	Nearest Neighbor	32
PR	Precision-Recall	54
PSW	Pre Swing	9

RF	Random Forest	17
RGB	Red Green Blue	3
RNN	Recurrent Neural Network	7
SDMM	Sub-action Depth Motion Map	143
SSTI	Skeleton Spatio-Temporal Image	130
STIPs	Spatio-Temporal Interest Points	6
SVM	Support Vector Machine	16
TCN	Temporal Convolutional Network	17
TP	True Positive	53
TST	Terminal Stance	9
TSW	Terminal Swing	8
UPCV1	University of Patras Computer Vision-1	88
UPCV2	University of Patras Computer Vision-2	88
UTD-MHAD	UTD Multimodal Human Action Dataset	134

List of Figures

1.1 Sample applications of HAR in smart environments. 3

1.2 Sample taxonomy of visual-data based HAR systems. 4

1.3 Sample frame sequence taken from (Shahroudy et al., 2016). (a) and (b) RGB. (c) and (d) Depth+Skeleton. (e) and (f) RGB+Skeleton. . . 5

1.4 Gait cycle phases, events, and their timings. 9

1.5 Thesis organization with respect to research objectives and contributions. 12

3.1 Proposed architecture for student's action recognition using RGB data. 41

3.2 YOLOv3 feature extractor (Darknet-53) (Redmon and Farhadi, 2018) 44

3.3 Sample of original images used in STUDENT_ACTION dataset. . . 48

3.4 Sample of augmented images in STUDENT_ACTION dataset. . . . 49

3.5 Action sample distribution in one of the fold in $10 - fold - Cross - Validation$ 50

3.6 Sample input and the corresponding output images. 51

3.7 Information about detected actions. 53

3.8 Precision/recall curves obtained for different actions. 55

3.9 mAP obtained during testing. 56

4.1 Proposed framework for 3D human action recognition using skeleton data. 61

4.2 Original skeleton format and new representation. (a) Skeleton joint data from Microsoft Kinect v1. (b) New tree representation of skeleton joints. 62

4.3 Confusion matrix using evaluation protocol P-1 for the entire dataset. 67

4.4 Confusion matrices. (a) AS1. (b) AS2. (c) AS3. 68

5.1 The workflow of the proposed human identification system using 3D skeleton-based gait features and the LSTM model. 76

5.2 Skeleton joints captured by Kinect depth sensor. (a) 20 skeleton joints by Kinect v1. (b) 25 skeleton joints by v2. 77

5.3 Gait cycle extraction steps: (a) Original and smoothed ankle distance-vector (b) Peaks in smoothed ankle distance-vector. 79

5.4 Extracted features. (a) BpLF. (b) JDF. (c) JAF. (d) InJDF. 82

5.5 Proposed LSTM based deep learning model. 84

5.6 Structure of LSTM unit. 85

5.7 Structure of GRU unit. 86

5.8 Rank-1 recognition performance of the proposed HI system using $Leave-one-sequence-out$ in K folds. (a) KGBD. (b) UPCV1. (c) UPCV2. 93

5.9 The CMC curves obtained using $Leave-one-sequence-out$ evaluation protocol. (a) KGBD. (b) UPCV1. (c) UPCV2. 94

5.10 The CMC curves of the proposed system using $Five-fold-Cross-Validation$. (a) KGBD. (b) UPCV1. (c) UPCV2. 97

5.11 CMC test curve obtained in $Random-split$ method on KS20. . . . 101

5.12 Comparison of total F-Value of feature groups in gait events. (a) UPCV1. (b) KGBD (continued in the next page). 103

5.13 Proposed human identification system using gait event-specific features and residual DL model with Attention. 108

5.14 Features extraction. (a) MLB. (b) DistJ. (c) MJDLR. 110

5.15 Proposed Attention guided residual LSTM. 111

5.16 CMC curves obtained. (a) KGBD. (b) UPCV1. (c) UPCV2 (continued in the next page). 123

6.1 Proposed multi-stream multi-attention deep learning model for HAR. 129

6.2 Sample DSTI and SSTI of actions. (a) MSRAction3D. [Top: DSTI, Bottom: SSTI] (b) UTD-MHAD. [Top: DSTI, Bottom: SSTI] . . . 132

6.3 Confusion matrix obtained for MSRAction3D. (a) AS1. (b) AS2. (c) AS3. 137

6.4 Confusion matrix obtained for UTD-MHAD. 138

6.5 Work flow architecture of the proposed HAR system. 140

6.6 Arranging sequence of frames into overlapping sub-actions. 142

6.7 Human skeleton joints considered and five regions 142

6.8 Sample MS-EDMM images generated for actions from UTD-MHAD dataset. (a) Boxing. (b) Sit to stand. (c) Stand to sit. (d) Lunge. . . . 144

6.9 Sample MS-EDMM images generated for actions from NTU RGB+D dataset: (a) Hand waving. (b) Rub two hands. (c) Take off a shoe. (d) Cheer up 145

6.10 Skeleton data processing stream (S1). 146

6.11 Depth data processing stream (S2). 147
6.12 Proposed residual LSTM with Attention block. 148
6.13 CONV block in proposed depth data processing stream. 150
6.14 Confusion matrix obtained for $C - S$ protocol on NTU RGB+D . . 156
6.15 Confusion matrix obtained for $C - V$ protocol on NTU RGB+D. . . 157
6.16 Confusion matrix obtained for $C - S$ protocol on UTD-MHAD. . . 158
6.17 Proposed multi-stream CNN-based deep learning model for HI. . . 163
6.18 Sample CSGEI images generated for gait cycles from IAS_Lab gait dataset: (a) Person-1. (b) Person-2. 164
6.19 Sample EDMM images generated for gait cycles from IAS_Lab gait dataset: (a) Person-1. (b) Person-2. 164

List of Tables

1.1 Advantages and disadvantages of different visual data modalities. . . 6

2.1 Summary of key existing HAR works based on RGB data. 17
2.2 Summary of key existing HAR works based on skeleton data. 21
2.3 Summary of key existing HAR works based on depth data. 24
2.4 Summary of key existing HAR works based on multiple data modalities. 26
2.5 Datasets for Human Action Recognition. 27
2.6 Actions in MSRAction3D. 28
2.7 Actions in UTD-MHAD. 29
2.8 Actions in NTU RGB+D. 30
2.9 Summary of key existing HI works based on data from Kinect depth sensors. 32
2.10 Summary of skeleton-based benchmark gait datasets. 35

3.1 Data augmentation techniques applied in the dataset. 42
3.2 Actions in RGB image-based STUDENT_ACTION dataset. 47
3.3 Distribution of actions among the frames. 47
3.4 Interpretation of Kappa statistic 52
3.5 TP and FP found for action labels while testing. 53
3.6 Comparison of video processing with and without template matching. 57

4.1 Three groups of actions AS1, AS2, and AS3. 64
4.2 Data sample distribution among testing and training in P-1 and P-2. 65
4.3 Action recognition Accuracy of the proposed model for HAR using skeleton data. 66
4.4 $Precision$, $Recall$ and $F - Score$ obtained for HAR using skeleton data. 69
4.5 Comparison of Accuracy (in %) of the proposed work with baseline methods. 70

5.1 Gait cycle events and duration. 74

5.2 Rank-1 recognition performance (in %) of the proposed HI system using $leave-one-sequence-out$ in various experimental setups (Best results are in bold). 91

5.3 Rank-1 recognition performance of the proposed HI system with different feature combinations using $Five-fold-cross-validation$ (in %) (Best results are in bold). 96

5.4 Comparison of Rank-1 $Accuracy$ of the proposed HI system using $Leave-one-sequence-out$ protocol with existing methods on KGBD. (Best result is in bold). 96

5.5 Comparison of the proposed HI system's Rank-1 $Accuracy$ using $Leave-one-sequence-out$ to existing methods on UPCV1. (Best result is in bold). 98

5.6 Comparison of the proposed HI system's Rank-1 $Accuracy$ using $Leave-one-sequence-out$ to existing methods on UPCV2. (Best result is in bold). 98

5.7 Comparison of the proposed HI system's Rank-1 $Accuracy$ using $Five-fold-cross-validation$ to state-of-the-art-works on KGDB. (Best result is in bold). 99

5.8 Comparison of the proposed HI system's Rank-1 $Accuracy$ using $Five-fold-cross-validation$ to existing methods on UPCV1. (Best result is in bold). 99

5.9 Rank-1 recognition performance of the proposed HI system on KS20 using $Random-split$ protocol (in %) (Best result is in bold). . . . 100

5.10 Comparison of Rank-1 $Accuracy$ to state-of-the-art works on KS20 dataset using $Random-split$ protocol. (Best result is in bold). . . 101

5.11 $Accuracy$ (in %) of proposed HI system on KS20 dataset using $Cross-view-split$ protocol and comparison with existing methods. . . . 102

5.12 Ablation experiments results. 102

5.13 Statistics of skeleton-based gait dataset used in the proposed work. . 113

5.14 Performance of proposed residual deep learning model on KGBD, UPCV1, and UPCV2 based on $Five-fold-cross-validation$. 114

5.15 Performance of proposed residual deep learning model on KGBD and UPCV2 using $Leave-one-sequence-out$ protocol. 115

5.16 Performance of proposed residual deep learning model on KS20 using $Random-split$ protocol, and on IAS-Lab the testing results on 'TestingA' and 'TestingB' 116

5.17 Comparison of Rank-1 $Accuracy$ (%) with state-of-the-art-works on $Five-fold-cross-validation.$ 117

5.18 Comparison of Rank-1 $Accuracy$ (%) of proposed work with state-of-the-art works: on IAS-Lab: two testing sets 'TestingA' and 'TestingB', on KS20: $Random-split$, on KGBD and UPCV2: $Leave-one-sequence-out$ 118

5.19 Rank-1 $Accuracy$ (%) on KS20 using $Cross-view-spilt$ and comparison with state-of-the-art works. 119

5.20 Rank-1 $Accuracy$ (%) of the different residual DL models without Attention module using $Five-fold-cross-validation.$ 119

5.21 Rank-1 $Accuracy$ (%) of the different residual deep learning model on KS20 using $Cross-view-split$ protocol without Attention module. 119

5.22 Rank-1 $Accuracy$ (%) of proposed residual DL model on KS20 without data augmentation. 120

5.23 Rank-1 $Accuracy$ (%) using different feature combinations using $Five-fold-cross-validation$ 121

5.24 Rank-1 $Accuracy$ (%) on different feature combinations on KS20 using $Cross-view-Split$ 122

6.1 Score fusion operations. 134

6.2 Recognition $Accuracy$ (in %) of the proposed system on MSRAction3D and UTD-MHAD. 136

6.3 Proposed system comparison with existing methods on MSRAction3D. 138

6.4 Proposed system comparison with existing methods on UTD-MHAD. 139

6.5 Score fusion operations. 150

6.6 HAR $Accuracy$ (in %) of proposed DL model and results of score fusion operations on UTD-MHAD dataset using $C-S$ protocol. [TA indicated Temporal Attention, Best results are in bold] 153

6.7 HAR $Accuracy$ (in %) of proposed DL model and results of score fusion operations on NTU RGB+D dataset using $C-S$ protocol. [TA indicates Temporal Attention, Best results are in bold] 154

6.8 HAR $Accuracy$ (in %) of proposed DL model and results of score fusion operations on NTU RGB+D dataset using $C-V$ protocol. [TA indicates Temporal Attention, Best results are in bold] 155

6.9 Top and bottom five $F-Score$ (in %) and corresponding actions recognized from NTU-RGB+D. 159

6.10 Top and bottom five $F - Score$ (in %) and corresponding actions recognized from UTD-MHAD. 160

6.11 $Accuracy$ (in %) of depth and fusion stream without Spatial Attention 160

6.12 Comparison of proposed work's recognition $Accuracy$ (in %) on UTD-MHAD with existing approaches 161

6.13 Comparison of proposed work's recognition $Accuracy$ (in %) on NTU RGB+D with existing approaches 162

6.14 $Accuracy$ (in %) of proposed multi-modal approach for HI and results of score fusion operations on IAS_Lab dataset. (Best results are in bold) 165

Chapter 1

Introduction

A smart environment can securely integrate multiple technological solutions to manage its assets, such as schools, transportation systems, hospitals, and community services. The smart environment contains contributions from eminent researchers describing techniques and issues related to its development and use in everyday life. It includes numerous components such as smart buildings, smart campuses, smart cities, smart classrooms, smart parking lots, smart transportation systems, etc. They utilize low-power sensors, cameras, and software with artificial intelligence to monitor the system's operation continuously. Proper monitoring technologies are needed to provide the secure living conditions and efficient management of resources in smart environments. The role of Computer Vision (CV) in a smart environment is very significant as it serves as the 'eyes' of the smart environment.

Several application services, such as parking systems, energy management, traffic management, health monitoring, waste management, transportation, etc., must be automated to build a smart environment. Therefore, numerous researchers are developing cutting-edge technologies to manage smart applications. Smart surveillance is one of the integral components of smart environment applications.

Recently, the proliferation of cameras used to monitor human activities has resulted in a massive amount of visual data. These cameras capture images in various formats. It requires considerable effort to interpret and store all of this visual data. Numerous applications rely on captured video content, including intelligent video surveillance to monitor human activity, crime detection, intelligent traffic management, etc. The primary goal of video surveillance is to observe a scene and search for specific human behaviour or action and incidences that may indicate emergence. Identifying individuals via video/images is also critical for maintaining a safe living environment. The traditional visual data analysis and content comprehension method require human resources, which adds high cost and time to the process. It is generally accepted that viewing video feeds requires a higher level of visual attention than the majority of daily activities. Specifically, the capacity to maintain attention and respond to infrequently occurring events is extremely demanding and error-prone due to attention lapses (Hampapur et al., 2003). As a result, sophisticated techniques for analyzing the visual data captured by the cameras are required. As the demand for security, improved living conditions, and optimal resource utilization increases in smart environments, the challenges associated with vision-based Human Activity/Action Recognition (HAR) and Human Identification (HI) increase proportionately. Hence, thesis work mainly focus on two important aspects of smart surveillance systems, namely: HAR, and HI.

1.1 Overview of Human Activity/Action Recognition

Human action is the movement pattern of various body parts, physically manifesting the individual's intentions and thoughts (Ramanathan et al., 2014). It is a collection of movements of human body parts with a particular semantic meaning (Liang et al., 2018). Examples of activities include walking, running, eating, hand waving, using a keyboard, clapping, falling, drinking water, jumping, fighting, etc. HAR is a process of recognizing the actions by analyzing the image sequence. The traditional way of action recognition involves manually analyzing the video sequences captured by the cameras, which is not only time-consuming but also needs more human resources. So, researchers started working on developing a system to learn about action patterns from the videos and use the knowledge gained to recognize the similar actions in other videos.

HAR is one of the challenging research topics which integrates computer vision, machine learning, pattern recognition, human detection in image/video, human pose estimation, and human tracking. In recent years it has grabbed the attention of researchers from academia, industry, and security agencies as it plays a key role in a wide range of smart environment applications such as smart video surveillance and smart home monitoring (Aggarwal and Ryoo, 2011; Ziaeefard and Bergevin, 2015), Human-Computer Interaction (Pickering et al., 2007; Papadopoulos et al., 2014), video indexing, and retrieval (Jan C. van Gemert and Snoek., 2015; Ramezani and Yaghmaee, 2016), so on. Figure 1.2 illustrates some of the applications of HAR in smart environments.

1.1.1 Categories of Human Actions

Human actions can be categorized into different groups (Vrigkas et al., 2015) depending on the complexity. For instance, "Gestures" are considered to be primitive movements of a person's body parts that may correspond to a specific action performed by that person. "Atomic actions" describe a specific person's motion that may be a part of more complex activities. "Human-to-object or human-to-human interactions" are actions that involve two or more people or objects. "Group actions" are actions performed by a group of people. "Human behaviors" are the outward manifestations of an individual's inner emotions, personality, and mental state. Finally, events are broad actions that

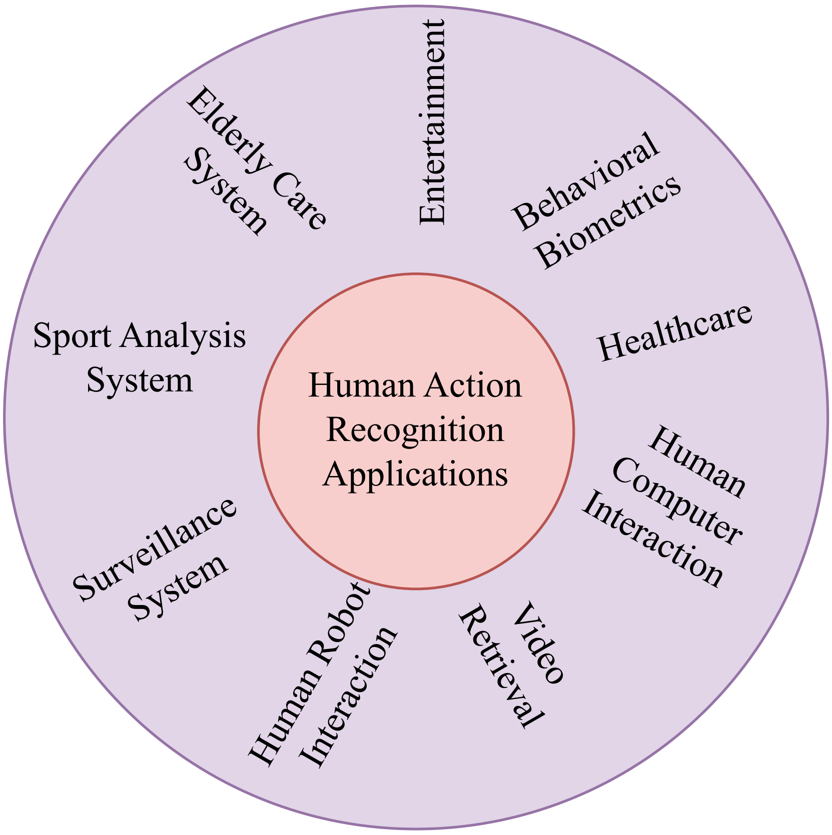


Figure 1.1: Sample applications of HAR in smart environments.

characterize interactions between people and reveal something about their goals or roles in society.

1.1.2 Categories of HAR Systems

The visual-data-based HAR systems can be categorized in different ways. Figure 1.2 shows the sample taxonomy of HAR systems. For example, depending on the number of images used, it can be a still-image based or video-based HAR system. Still-image based HAR system uses a single image for action recognition. For example, simple human actions like gestures and atomic actions can be recognized with this. Even though it is cost-effective, all the actions cannot be recognized with a single image. There is a need to consider the movement in the sequence of image frames. The video-based HAR system uses a sequence of frames for interpreting the human action in the scene. This extracts spatial and temporal characteristics from the image sequence to perform the classification.

Another way of categorizing the HAR system is depending on the visual data modality. The visual data modalities can be Red Green Blue (RGB), skeleton, depth, infrared,

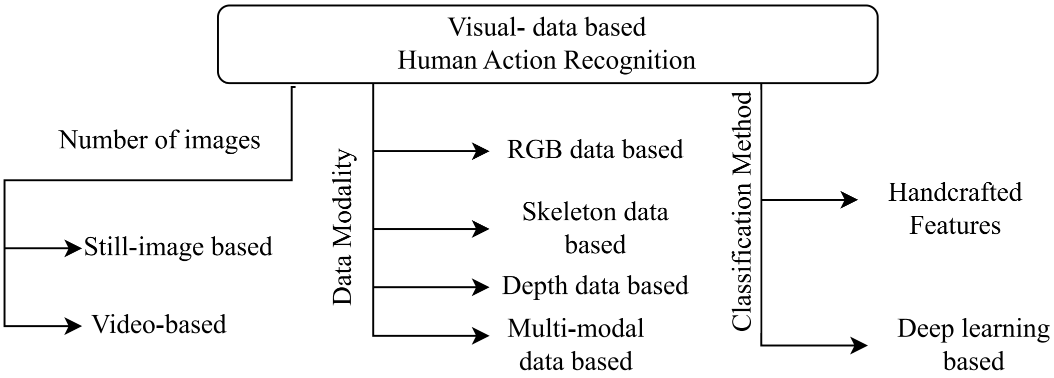
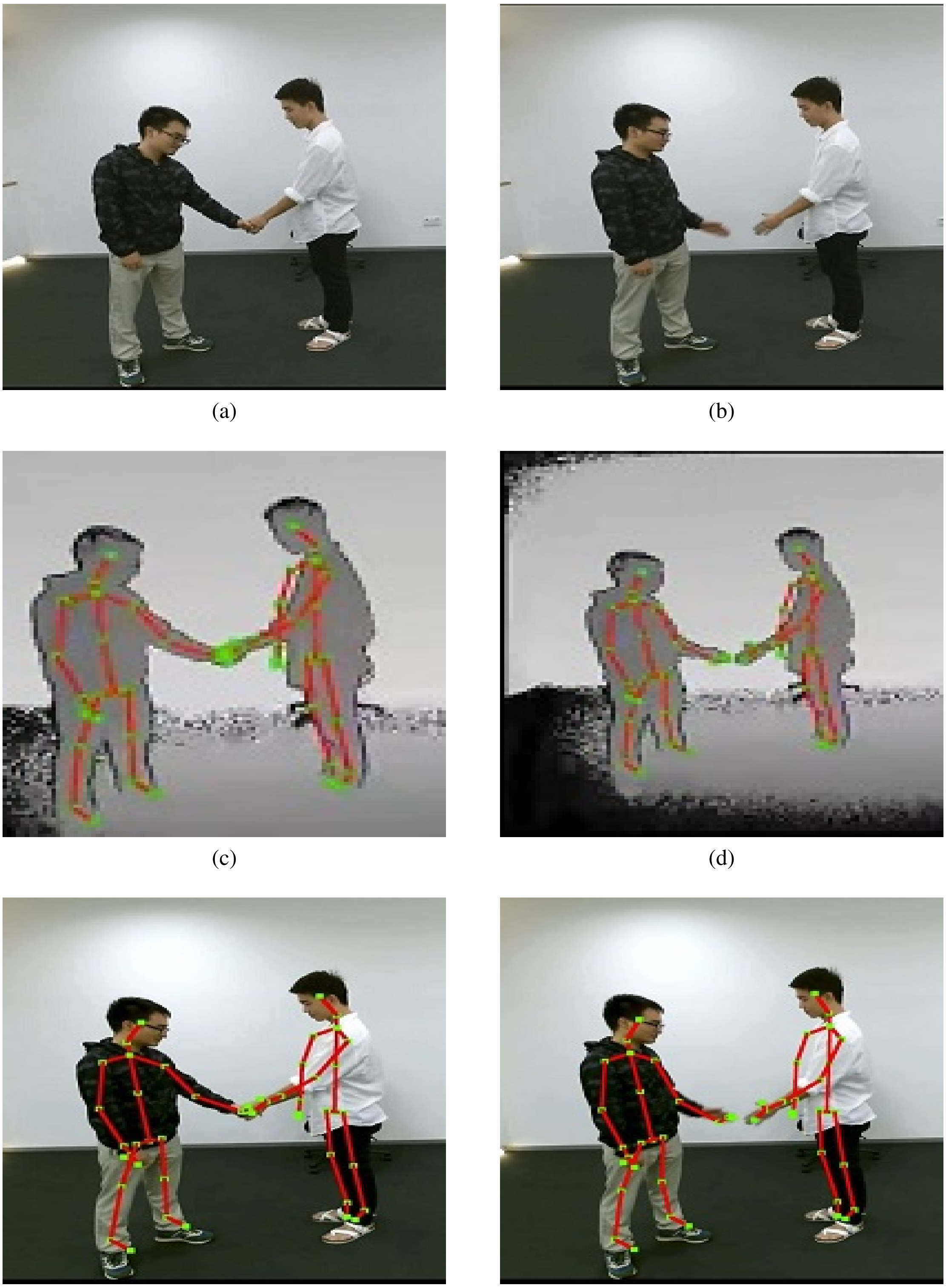


Figure 1.2: Sample taxonomy of visual-data based HAR systems.

and pointcolud etc. The most common types of data used by the HAR system are RGB, skeleton, and depth. There are several uni-modal HAR systems which use any one of the said data modalities for HAR. Whereas, multi-modal systems use combination of data modalities (Wang et al., 2020; Sun et al., 2023).

The RGB images recreate what the human eye sees and provide values for the red, green, and blue components. They provide detailed information about the appearance of the captured scene's context. Skeleton data provides the coordinate positions of different body joints. The skeleton joints can be derived from RGB or depth images by applying pose estimation algorithms. A vast number of skeleton data-based HAR systems are based on Three Dimension (3D) skeleton data provided by Kinect depth sensor. Depth maps are images in which the pixel values depict the distance between a given viewpoint and the scene's points. The depth modality, which is frequently insensitive to variations in colour and texture, provides accurate 3D structural and geometric shape information of human subjects and can thus be used for HAR. The sample frames from NTU RGB+D multi-modal human action dataset (Shahroudy et al., 2016) are shown in Figure 1.3. The first row shows two frames from RGB video. Second row depicts the depth and skeleton joints. RGB and skeleton data is shown in third row.

Based on the approach used for classification the HAR system can be either handcrafted feature based or Deep Learning (DL) model based. Handcrafted features for action recognition are the features derived using some algorithms, from the information available in videos or images. They capture human body movements, spatial and temporal changes in the action video, and mainly used in machine learning algorithms for action classification. Examples for handcrafted feature-based action representation are image sequence based representation, the trajectory of skeleton joints based representation, etc.



(a) (b) (c) (d) (e) (f)

Figure 1.3: Sample frame sequence taken from (Shahroudy et al., 2016). (a) and (b) RGB. (c) and (d) Depth+Skeleton. (e) and (f) RGB+Skeleton.

In the fixed camera scenario, the background subtraction technique is used to determine the shape information from the RGB data. Spatio-Temporal Interest Points (STIPs) detection is used to determine regions that have movement change and to represent action in the video (Laptev and Lindeberg, 2005). Some approaches are based on capturing the trajectories (Wang et al., 2011). HAR, based on depth sequence, uses changes in the depth map to represent action as work in (Yang and Tian, 2014; Rahmani et al., 2014; Yang and Tian, 2017). The skeleton data-based HAR systems can be either joint based or body part based methods (Li et al., 2018; Vemulapalli and Chellappa, 2016; Vemulapalli et al., 2014). They create feature vectors based on skeleton joints or builds a human model from skeleton data and extracts features for classification. All these modalities have several advantages and disadvantages. So, works are found by combining the different modalities to compensate the shortcomings (Sun et al., 2023). Table 1.1 explains the pros and cons of different data modalities.

Table 1.1: Advantages and disadvantages of different visual data modalities.

Modality	**Advantages**	**Disadvantages**
RGB	• Provides the detailed visual information about the surrounding environment. • Easy to collect and operate. • Huge number of applications are based on RGB data.	• Sensitive to viewpoint, background, and lighting condition of the scene. • Due to large data size, it demands for high computational cost, more resources.
Skeleton	• Provides the 3D structural data regarding subject pose. • Most informative and easy to use. • Tolerant to viewpoint, background, color of cloth. • Demands less computation and resources.	• Absence of of appearance, shape information. • Noise.
Depth	• Provide 3D structural and geometric shape information. • Noise.	• Lack of color and texture information. • Restricted working distance.

The use of DL models in computer vision is dramatically increasing due to the superior performance in various classification tasks. So, building advanced DL models to improve the HAR system is alluring to the researchers. The most widely used DL concepts in HAR using different data modalities are Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and its variants. The CNN has shown its superior performance in learning the image features for the classification task. So, it is the most widely used method in RGB and depth data-based approaches. In skeleton data-based approaches, often the features from skeleton data are represented as images, and classification is done using CNN models. Since RNN and its variants proved that they are efficient in capturing temporal features for time series tasks, researchers explored using Long Short-Term Memory (LSTM), RNN, and Gated Recurrent Units (GRU) for video-based HAR.

1.2 Overview of Gait-based Human Identification

The number of cameras deployed to monitor people in today's smart environments like airports, shopping malls, university campuses, and workplaces has increased rapidly in recent years in response to rising crime and terrorist threats. Finding a person from these videos through manual analysis is time-consuming and laborious. Because of human factors like fatigue and boredom from keeping watch for long periods of time, lack of interest, and distractions, there is a high chance of mistakes being made due to the sheer volume of monitors that need to be checked. So, there is a demand for accurate, remote human identification in a variety of embedded applications within intelligent surveillance systems is rising dramatically in smart environments. A wide variety of real-world applications benefit greatly from the use of such systems, including forensics, monitoring for terrorist activity, crime prevention, and access control (Huynh-The et al., 2020). In these contexts, the ability to quickly and accurately identify a single person out of a large group using only their biometric characteristics is of critical importance.

There are several existing approaches for HI, using several type of biometrics such as fingerprint, face, iris scan, fingerprint, and voice etc. Due to limited field of view and other factors, identifying individuals in video surveillance systems is difficult. Surveillance systems that rely on human identification through face recognition are ineffective when the face is hidden by a mask, hand, or hat (Batchuluun et al., 2018; Sepas-Moghaddam and Etemad, 2023). Gait-based human identification is currently a thriving area of study as a means of surmounting these difficulties. Unlike these traditional approaches, which require human attention for identification, the gait based human identification is non-intrusive. Human gait refers to locomotion achieved through the cyclic movement of human limbs (Boyd and Little, 2005). In gait-based human identification, both data collection and classification is done without the subject's knowledge (Khamsemanan et al., 2018). Gait-based human recognition is an emerging behavioral biometric trait for intelligent surveillance monitoring because of its non-contact and non-cooperation with subjects (Singh et al., 2018). Research on gait has drawn the attention of several researchers due to the following key benefits (Tafazzoli and Safabakhsh, 2010; Khamsemanan et al., 2018; Singh et al., 2018; Verlekar et al., 2018):

- Each individual has a unique gait.
- Gait of a person is impossible to hide.
- Imitating of another person's walking style is not possible.
- Its unobtrusive nature, i.e., gait data are collected at a distance without the subject's knowledge.
- It is much more difficult to continuously alter a person's gait characteristics.
- Low-resolution video sequences can be analyzed for gait characteristics.
- Gait recognition still works well while features such as face images are hidden.

1.2.1 Gait Cycle Components

According to (Kastaniotis et al., 2016), the entire gait cycle consists of two phases: Stance and Swing, with any leg serving as a reference point. The entire time that a foot is on the ground constitutes the stance phase. The duration that the foot is in the air is called swing phase. 60% of the gait cycle consists of the Stance phase, while the remaining 40% consists of the Swing phase. These phases are subdivided into multiple events. A gait cycle is formed by heel strike 'Initial Contact (IC)' of a leg to the floor to subsequent heel strike 'Terminal Swing (TSW)' of the same leg. Consequently, a gait cycle is composed of three successive heel strikes. During heel strike, the space between person's ankles will be greatest. Hence, based on peaks in ankle distances the gait cycles in walking sequence are detected. Figure 1.4 displays the various events and their relative proportion during a gait cycle. We adapted the timing distribution of the gait cycle's various events based on the information found in (Webster and Darter, 2019). Stance phase of gait comprised of four events: Loading Response (LR), Mid Stance (MST), Terminal Stance (TST), Pre Swing (PSW). Swing phase comprised of three events: Initial Swing (ISW), Mid Swing (MSW), and TSW. Each of these have fixed duration. All these events are defined by position of foot of person during walking.

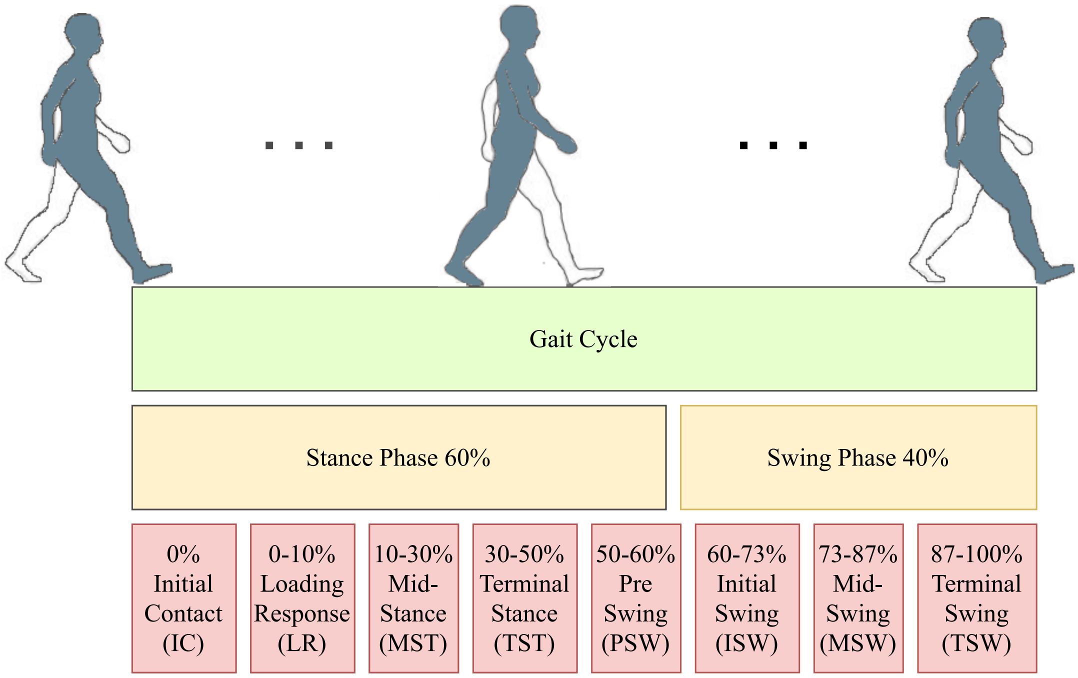


Figure 1.4: Gait cycle phases, events, and their timings.

1.3 Approaches for Gait Recognition

There are mainly two broad categories of gait recognition, namely: model-free, and model-based (Chai et al., 2011; Singh et al., 2018; Rida et al., 2019).

Model-free Approaches

Features are extracted directly from the part of gait contour (Singh et al., 2018). These approaches do not need the prior knowledge of the model. The majority of these methods based on silhouettes images of a person. Where, silhouette of a person is extracted by background subtraction method.

Model-based Approaches

A sequence of static or dynamic body parameters are obtained via modeling or tracking body components such as limbs, legs, arms, etc., (Chai et al., 2011; Mahfouf et al., 2018). In this a prior model is established to match real images. The gait features like length of stride, angular measurement, anthropometric features, joint trajectories are extracted from the structured body model by fitting the model to observed body of the person. Most specifically human skeleton data and body structures form the basis for classification in model-based gait recognition methods (Khamsemanan et al., 2018). The work on model-based gait recognition is proliferated with the development of the Microsoft Kinect depth sensors. The 3D skeleton data information provided by the Kinect depth sensor eliminated the complex algorithm of building human model from Two Dimension (2D) images. This thesis work focuses on HI using skeleton data based gait features.

1.4 Motivation

Vision-based HAR and HI have the potential to play a crucial role in a variety of smart environment applications, including smart buildings, smart cities, and smart campuses. Global security threats have generated a substantial demand for intelligent surveillance systems in smart environments to provide a secure living environment and efficient management through appropriate monitoring technologies. Several surveillance cameras are installed in smart environments to monitor people and maintain a safe and comfortable living environment. These cameras generate vast quantities of video data, which is manually analyzed in conventional surveillance. Due to humans' limited patience, manually analyzing vast amounts of video data is time-consuming, labor-intensive, and prone to failure. Therefore, automated video analysis systems are required for smart environments to provide a better living environment. Thus there is a demand for efficient HAR, and HI systems to support various applications of smart environments.

The actions of humans vary in different application domains. Consequently, there is a significant demand for developing application domain-specific action datasets and real-time action recognition systems. In addition, an unobtrusive HI system is required to provide a secure environment in which biometric data can be collected and processed without the individual's knowledge. Gait is one of the biometrics that can be remotely collected without the subject's cooperation. Therefore, gait-based human identification systems can potentially play a role in the forensics department of smart environments. The advancement in camera technologies allows video data in different modalities with pros and cons. As a result, the analysis of diverse video data in various modalities for HAR and HI is gaining importance in the present day.

Human actions are diverse in nature. Different individuals can perform the same action at different speeds. In addition, the movement of a person's limbs during action will vary slightly among individuals. Also, the same action appears differently from multiple viewpoints. Various actions can be remarkably similar to one another. All of these issues contribute to the poor performance of HAR systems. Consequently, there is a demand for sophisticated HAR systems to improve system performance by addressing these issues.

Human gait recognition is one of the most difficult and alluring fields in surveillance applications. The more advanced camera technology provides data in different data modalities. The advantages of 3D data, particularly skeleton data, are numerous. To improve the performance of HI systems, work on skeleton data-based gait recognition is therefore thriving.

1.5 Organization of Thesis

Figure 1.5 illustrates the details about the thesis organization including formulated research objectives and the respective research contribution chapters. The remaining part of this thesis is organized as follows.

- Chapter 2 - **Literature Review**
  This chapter reviews the related work HAR systems with respect to different modality of data, the approach of feature extraction and classification. In addition, the HI systems based on gait biometric using 2D and 3D data. Specific emphasis is placed on gait recognition based on skeleton data. In addition, the benchmark 3D datasets for HAR and skeleton-based gait recognition are described. Following the outcomes of the literature review, the problem statement and research objectives are framed.

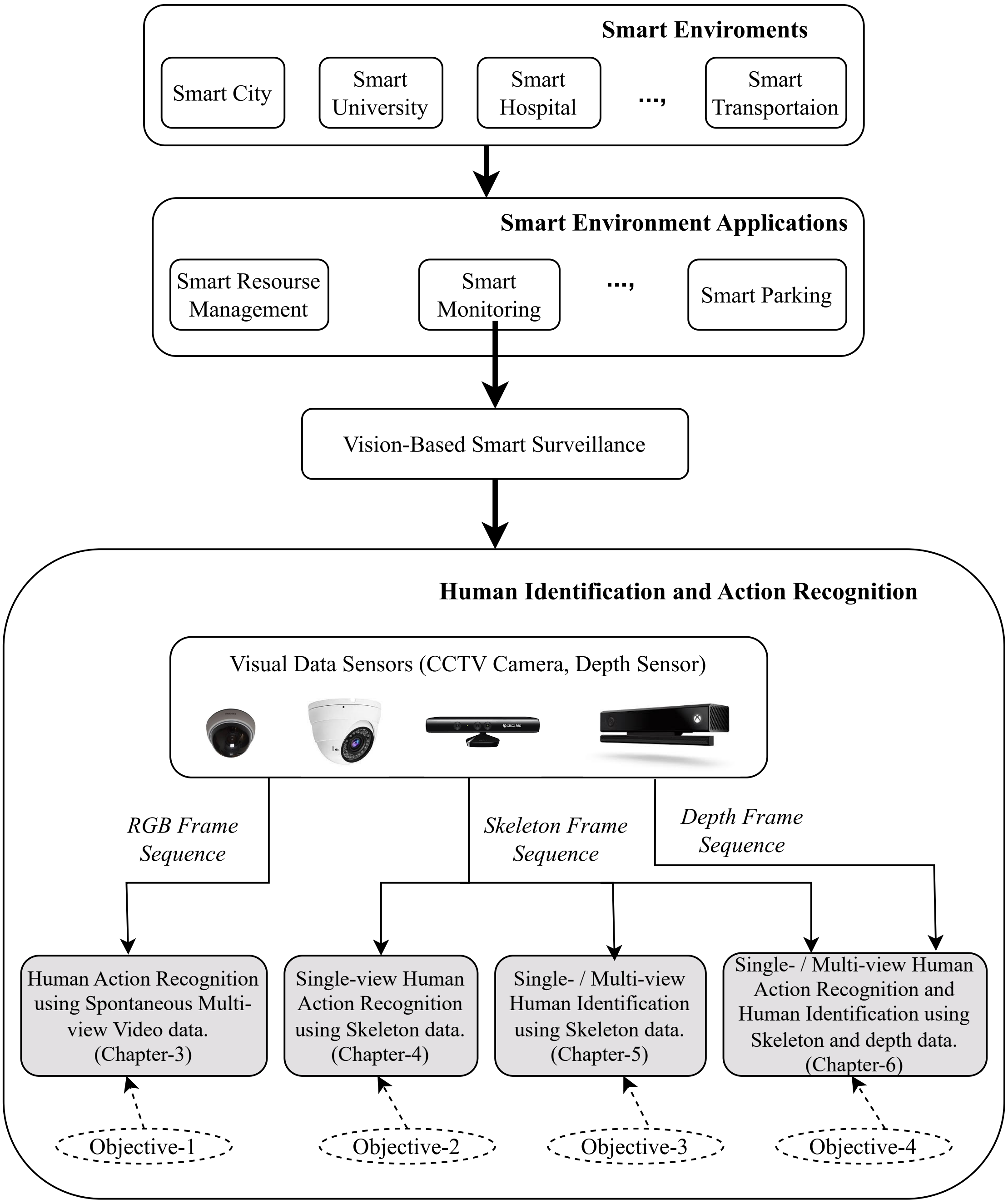

Smart Enviroments
Smart City
Smart University
Smart Hospital
...,
Smart Transportaion
Smart Environment Applications
Smart Resourse Management
Smart Monitoring
...,
Smart Parking
Vision-Based Smart Surveillance
Human Identification and Action Recognition
Visual Data Sensors (CCTV Camera, Depth Sensor)
RGB Frame Sequence
Skeleton Frame Sequence
Depth Frame Sequence
Human Action Recognition using Spontaneous Multi-view Video data. (Chapter-3)
Single-view Human Action Recognition using Skeleton data. (Chapter-4)
Single- / Multi-view Human Identification using Skeleton data. (Chapter-5)
Single- / Multi-view Human Action Recognition and Human Identification using Skeleton and depth data. (Chapter-6)
Objective-1
Objective-2
Objective-3
Objective-4

Figure 1.5: Thesis organization with respect to research objectives and contributions.

- Chapter 3 - **Domain Specific HAR using RGB Data**
  This chapter provides a comprehensive description of a context-specific HAR system utilizing RGB data for student action recognition in smart computer labs. It provides information about the creation of a dataset, which is the action defined for the said domain and also the performance of the proposed system in action recognition.

- Chapter 4 - **HAR using Skeleton Data**
  This chapter describes a HAR system for single-view action recognition using 3D skeleton data. This elaborates on a new tree representation of skeleton joints, and traversal of nodes based on Depth First Search (DFS) method for feature extraction. Also provides information about the performance on one of the challenging dataset using DL model.

- Chapter 5 - **HI using Skeleton Data**
  This chapter discusses the two different approaches for HI based on skeleton-based gait data. Here, novel gait-event-specific quantitative summaries of various sets of features are described. Also, the advanced DL models based on LSTM / GRU, and Attention units are discussed. Further, the various experiments using benchmark single and multi-view datasets with state-of-the-art protocols are discussed in detail.

- Chapter 6 - **HAR and HI using Fusion of Skeleton and Depth Data**
  This chapter discusses the approaches for multi-modal HAR and HI using a combination of skeleton and depth data, multi-stream DL models, and score fusion operations. Two distinct approaches to multi-modal HAR are presented here. The first work discusses the RGB image representation of action in skeleton data streams and the spatio-temporal single image representation of human action in depth frame sequences. In addition, a multi-stream DL model with temporal Attention is also described for learning features from these representations. The second work discusses the investigation of sub-actions of human actions for action representation and feature extraction from the skeleton and depth stream of the action. In addition, a multi-stream DL model with spatial and temporal Attention is discussed for learning the features. Also, both works elaborate on the various experiments utilizing single- and multi-view benchmark action datasets with standard evaluation protocols.

  This chapter also discusses a proposed HI system employing gait data in skeleton and depth format. The proposed HI system describes the spatio-temporal image representation of the gait cycle in skeleton and depth format. In addition, it describes a multi-stream DL model for training these images for gait recognition. Finally, this section on HI system elaborates on the experimental findings on a small-scale multi-modal gait dataset.

- Chapter 7 - **Conclusions and Future Directions**
  This chapter concludes the thesis by summarizing the thesis contributions and highlighting possible future directions for unobtrusive HAR and HI systems to support vision-based smart applications.

1.6 Summary

This chapter provided a detailed explanation of the requirements for vision-based intelligent surveillance systems. Concerning surveillance systems, the roles of HAR and HI are highlighted. The fundamental concepts of data modalities of human actions are elaborated upon. Also discussed are the various approaches to categorizing the HAR. In addition, the fundamental concepts of human gait are presented. Furthermore, the various approaches to gait recognition are discussed. In addition, the motivation for this research and the thesis structure are presented. The subsequent chapter discusses the comprehensive literature review, identified research gaps, problem statement, and research objectives.

Chapter 2

Literature Review

A comprehensive literature review was conducted to understand the work performed by researchers in vision-based HAR and unobtrusive HI for supporting vision-based smart applications in smart environments. In light of the potential benefits in various applications, vision-based HAR and HI have piqued the attention of researchers and smart environment application builders. In this section, we provide an in-depth explanation of the vast spectrum of significant existing works on vision-based HAR, HI systems, the data modalities, and datasets for vision-based HAR, HI that have led to the development of intelligent vision-based smart environment applications. We then discuss the literature review findings, followed by the problem statement and primary research objectives.

2.1 Human Action Recognition

The primary goal of HAR is to develop an automated system that would mimic the human visual system in understanding and describing the human actions in a given scene (Abu-Bakar, 2019). It consists of data acquisition from the sensor, pre-processing, segmentation, feature extraction, training, and classification (Khurana and Singh Kushwaha, 2018). The sensors utilized by HAR may be visual or non-visual. Based on visibility, data modalities can be roughly divided into two categories: visual and non-visual (Sun et al., 2023). As this thesis work focuses on strengthening surveillance systems for smart environments, hence we focus only on visual data modalities from visual sensors such as depth sensors and RGB cameras.

From the perspective of complexity of human actions the HAR can be categorized into still image-based HAR (Ko et al., 2015; Yan et al., 2017; Sreela and Idicula, 2018), and video-based HAR (Yang and Tian, 2017; Liu et al., 2018; Kamel et al., 2019) systems. Simple human actions, such as Drinking, Using Mobile, etc., can be identified from a single image frame. Recognizing a complex set of actions, such as walking, running, etc., requires a sequence of image frames with temporal features. Traditional spatio-temporal features cannot be used for action recognition in still images as they provide only spatial information. Researchers have sought out various high-level cues in still images to recognize actions better than using low-level features in the whole image. The human body, body parts, action-related objects, human-object interaction, and the entire scene or context are some of the most widely used high-level cues for still image-based action recognition (Guo and Lai, 2014). A complex action comprises of temporal features between frames can be recognized with video-based HAR systems. In video-based HAR the temporal features from set of frames along with spatial information is considered for accurate recognition of complex human actions.

Recently, the visual data can be captured and stored in different modalities. In general, the visual modalities such as RGB, skeleton, depth data of a scene are most "intuitive" for representing human actions. Accordingly from the perspective of data modalities various number of approaches are proposed in the literature using RGB data (Hoang et al., 2018; Angelini et al., 2019), skeleton data (Yang et al., 2019; Jiang et al., 2020; Shao et al., 2021; Li et al., 2022) , and depth data (Yang et al., 2012; Xiao et al., 2019). Also, nowadays HAR based on fusion of different data modalities is alluring researchers (Dhiman and Vishwakarma, 2020; Yang et al., 2020). Below we discuss the prominent contributions found in literature study for HAR using different visual data modalities such as RGB, skeleton, and depth.

2.1.1 RGB Image-based HAR

HAR based on RGB images is a popular research topic in computer vision and pattern recognition. Images and videos captured with RGB cameras that attempt to replicate what human eyes see are often referred to as being in the RGB modality. RGB data is typically simple to collect and contains substantial information regarding the scene context. HAR from RGB data is challenging due to varying backgrounds, viewpoints, scales of humans, and lighting conditions. Based on the number of frames used for HAR, there are video-based and still image-based HAR systems using RGB data.

Most of the earlier works on HAR from still images are RGB data based. Many spatio-temporal features and methods developed for traditional video-based action recognition are inapplicable to still images (Guo and Lai, 2014). However, the computation and resource requirements for still image-based HAR are significantly lower than those for video-based HAR, due to the reduced number of features that must be processed and stored. Numerous approaches (Delaitre et al., 2010; Yao and Fei-Fei, 2010; Yan et al., 2017; Sreela and Idicula, 2018) were proposed in the literature for recognizing human actions from still/static images. The work in (Delaitre et al., 2010) proposed action recognition in still images by combining bag-of-features methods and the part-based Latent Support Vector Machine (SVM). The method specifically investigates the role of background scene context in HAR. Eventually, it combines the statistical and part-based representations and the individual detail with the scene context for action recognition. (Rădulescu and Florea, 2021) presented and compared the performance of three deep learning models with 2D, 3D kernel, and Temporal Convolutional Network (TCN) units. The authors determined that a model with a 2D kernel is the quickest but has poor performance. Alternatively, TCN performed better. The summary of key existing works on RGB based HAR is reported in Table 2.1.

Table 2.1: Summary of key existing HAR works based on RGB data.

Authors	**Methodology**	**Remarks**
(Ko et al., 2015)	A two-layer classification model with poselet-based features for HAR from images. Two layers of Random Forest (RF) based classification are applied on the poselets to recognize actions.	Infers the relationship between context and poselets among human actions. Miss-classification in complex background, and unclear human poselet.
(Zhang et al., 2016)	Selective search is used to generate initial object proposals, which are then disassembled into finer-grained object parts for use in locating Human-Object Interaction (HOI) zones. Actions are predicted using HOIs.	Attempt to solve problems encountered by annotators. The HOI extraction for various actions requires improvement.
(Yan et al., 2017)	CNN features for image patches are generated using a region proposal algorithm, and are used to encode the image as a compact code that captures the image's fine-grained properties and global context.	Better action recognition from still images using local patches and global context.
(Sreela and Idicula, 2018)	CNN features of the image are extracted using a residual neural network and then classification using SVM classifier.	Achieved better performance in recognizing certain actions in the dataset. Must be enhanced to recognize all actions.

Table 2.1 Continued from previous page.

Authors	Methodology	Remarks
(Ullah et al., 2018)	Using CNN model, features are extracted from every sixth frame of an action video and then learned using Bi-directional LSTM (DB-LSTM) for HAR.	Capable of learning long term complex sequences in videos. Features from salient regions of the frame need to considered for action recognition.
(Li et al., 2018)	In this, the authors presented a framework to select the discriminative part in the spatial dimension and used multiple layers of a LSTM to learn temporal features for HAR.	Enriches spatio-temporal information of action by combining spatial and temporal features. More structural features are required for better performance.
(Gnouma et al., 2019)	Silhouette images are extracted by detecting the foreground. These images are combined to create a History of Binary Motion Image (HBMI), and trained with an Artificial Neural Network (ANN) for HAR.	Model is independent of the style of the individuals as binary foreground masks are utilized.
(Tu et al., 2019)	Aggregates action video features spatially and temporally by encoding deep features in both sub-actions spatially and action-stages temporally. It divides the local deep features into segments and chooses the informative features from each segment.	A method for accurately predicting the discriminative significance of each frame by ignoring repetitive, unimportant, or noisy frames that are less useful or even detrimental to the target action.
(Ji et al., 2020)	Extracted semantic contexts with interactive objects, scenes, and body motions from action videos to construct a context knowledge map, followed by classification.	Using four existing models to extract context information could result in high computational cost.

Many existing approaches to recognizing human actions in videos rely on classifiers applied to sequences of RGB colour images. While these approaches performed well in recognizing simple human actions like running and bending against a simple background, they are highly sensitive to factors influencing RGB image quality. The factors like complex backgrounds, illumination variation, and clothing colour, viewpoints, and human scales make it challenging to segment the human body in all scenes. Furthermore, each person has their own unique way of moving their body when accomplishing the same task semantically. However, if two separate actions share a similar motion path, it becomes more challenging to recognize them. Recognizing the action, especially when performed in the camera's direction, can be difficult if depth cues are absent in RGB images. In addition, RGB videos typically have large data sizes, resulting in high computational costs when modeling the spatiotemporal context for HAR (Sun et al., 2023). Recent human action recognition technologies have considered using depth cameras to provide 3D data in skeleton and depth data to overcome the limitations mentioned above.

2.1.2 HAR using Skeleton Data

The skeleton data provides the 2D/3D coordinates of various joints in human bodies for each frame of the action video, compared to the entire image frame. Thus, it drastically reduces the storage and computation requirements and consequently, the resource requirements. The trajectories of the body's joints, which are recorded in the skeleton sequences, provide the most useful information about human motion, making the data suitable for HAR. Skeleton data has many benefits for HAR, such as pose information, a simple and informative representation, scale in-variance, and resistance to variations in clothing textures and backgrounds. These benefits, combined with the accessibility of high-quality, low-cost depth sensors, have piqued the interest of the scientific community in skeleton-based HAR. Numerous works utilizing skeleton data, particularly 3D skeleton data, have been reported in the existing literature on HAR. Many of the earliest HAR research efforts concentrated on hand-crafted spatial and temporal features from skeleton sequences (Aggarwal and Xia, 2014; Zhang et al., 2016). In past few years because of its superior feature-learning ability, DL has quickly become the method of choice for skeleton-based HAR.

The handcrafted feature extraction may focus on joint based or body-part based features for HAR (Li et al., 2018). (Ofli et al., 2014) developed a method for selecting most informative skeleton joints for HAR. Selecting a key joint and then extracting features from the remaining joints in relation to that joint is the basis of the proposed methods in (Yang and Tian, 2012) and (Vemulapalli et al., 2014). A compact representation of postures was proposed by (Xia et al., 2012), which characterised human postures as histograms of 3D joint locations within a spherical coordinate system. In body-part based approaches, the human body parts are used to model the human's articulated system. These body parts are typically represented by rigid cylindrical shapes connected at their joints. (Vemulapalli and Chellappa, 2016) used relative 3D rotations between joints as feature for HAR. (Amor et al., 2016) suggested a comprehensive framework for action analysis with skeleton shape evolution.

The DL methods are based on RNN (and its gated variations like LSTM), CNN, and Graph Convolutional Network (GCN). RNN and its variations showed its superiority in learning the temporal dependencies in sequential data. Consequently, these have been applied and adapted in a number of different ways (Zhang et al., 2018; Liu et al., 2018) to effectively model the temporal context features within the skeleton sequences for HAR. (Liu et al., 2018) introduced DL model based on LSTM to analyze the 3D coordinates of skeleton joints at each frame and processing step. Furthermore, a skeleton tree traversal method that uses the adjacency graph of body joints can accurately capture the data and boost the performance. Using a global context memory cell, (Liu et al., 2018) focused attention to the most informative joints in human action for more accurate HAR. (Jiang et al., 2020) proposed a spatial-temporal skeleton transformation descriptor that is learned with an LSTM network.

CNNs have achieved great success in the field of 2D image analysis due to their superior ability to learn features in the spatial domain. However, when dealing with skeleton-based HAR, it becomes difficult to model spatio-temporal data. Numerous sophisticated methods have been suggested, such as using temporal convolution on skeleton data (Kim and Reiter, 2017) or representing skeleton sequences as images that are then fed to regular CNNs for HAR (Wang et al., 2018). (Hou et al., 2018) represented the spatio-temporal features in a sequence of skeleton frames as colour images , followed by CNN model to learn the features for action classification. (Wang et al., 2018) proposed encoding of joint trajectories and their dynamics into color images, and a CNN model to learn the discriminative features for HAR.

Recently, numerous methods for HAR based on GCN have been proposed, which treat skeleton information as edges and nodes in the graph (Ahmad et al., 2021). (Yan et al., 2018) proposed a spatial temporal GCN for HAR from skeleton data. The model constructs a set of spatial temporal graph convolutions on the skeleton sequences to extract motion information. (Zheng et al., 2022) used bone and joint streams of skeleton data and proposed a two-stream GCN for HAR from skeleton data.

Spatio-temporal graph routing to adaptively learn the intrinsic high-order connectivity relationships for physically apart skeleton joints is proposed in (Li et al., 2019). This consists of two components: a spatial graph router for tracing the connectivity relationships between the joints and a temporal graph router for analysing structural data. There are different approaches to tackle different issues related to skeleton based HAR. In this we primarily focus on DL models based on CNN and RNN variants for HAR from skeleton data. The summary of key existing works on skeleton data-based HAR which inspired this thesis work is given in Table 2.2.

Table 2.2: Summary of key existing HAR works based on skeleton data.

Authors	**Methodology**	**Remarks**
(Du et al., 2016)	The human skeleton is divided into five major regions based on the physical structure and fed to RNN-based DL models.	Appearance features can be combined with the temporal features to enhance the performance.
(Wang and Wang, 2018)	HAR is based on primitive geometries such as joints, edges, and surfaces of skeleton data. The RNN-based DL model with a viewpoint transformation layer was introduced for classification.	Shown the benefits of incorporating multiple geometric primitives. For enhanced performance, geometric relationships and temporal skeleton dynamics can be investigated.
(Ke et al., 2018)	Each channel of the 3D coordinates of a skeleton sequence is converted into a clip containing the spatial and temporal information of the series of skeletons. Also, Multitask Convolutional Neural Network (MTCNN) is proposed to learn these clips for HAR.	Considers both spatial and temporal features of HAR. Needs more computations as there are three parallel CNNs.

Table 2.2 Continued from previous page.

Authors	Methodology	Remarks
(Liu et al., 2019)	DL model is used to model the context and dependency information in spatio-temporal dimension.	Main contribution is a large scale dataset. Use of multi-modal features can increase recognition performance.
(Li et al., 2019)	Accurate HAR is achieved by constructing three views in the spatial domain using enhanced joint trajectory maps and feeding them to a stack of CNN, and LSTM networks to learn spatio-temporal data.	Uses temporal and spatial information together in learning the features for better performance.
(Pham et al., 2019)	Action is represented as a RGB image by dividing the skeleton into five parts, and normalizing the 3D coordinates. These are trained using CNN model for HAR.	Better performance at low computation power. Further evaluation is required on multi-view datasets.
(Agahian et al., 2020)	HAR framework uses pose representation and encoding. Defined a pose descriptor with normalized skeleton joint coordinates, displacement information relative to temporal offset, and previous timestamp.	Considers all joints to generate features. Focus on most discriminative features improves the performance.
(Shao et al., 2021)	Proposed a new perspective on view-invariant action recognition from skeleton data. The descriptors of actions are derived from their skeleton self-similarities and trained with a multi-stream DL model.	Multiple CONVNets and LSTM captures the spatio-temporal information from multiple views leads to good performance, with more computation requirement.
(Ding et al., 2021)	The dynamic skeleton data from a series of frames are represented in a 3D grid structure capturing the inter-dependencies between body parts. A dual-stream 3D CNN model is used to learn the features for HAR.	Model demands for more computational resources. The action representation could be improved, which would result in improved performance.

Table 2.2 Continued from previous page.

Authors	Methodology	Remarks
(Sameem et al., 2021)	Proposed an action descriptor based on 3D skeleton joint's angle, distance and frame-frame interrelationships, followed by classification.	Can investigate DL models to enhance performance. Experiments on view-invariant HAR are required.
(Liu et al., 2022)	The action in a sequence of skeletons is converted into a spatial-temporal graph structure and an image depicting skeleton motion. Followed by a GCN to learn the action representation.	Performed feature fusion of dual streams. The score fusion of different streams from multi-view features may further improve the performance.
(Ng et al., 2022)	Each skeleton is initially divided into different groups. After that, the auto-encoder-based DL is proposed for HAR. The purpose of the attention mechanism is to concentrate on more informative body parts.	Performance need to be improved by exploring different Attention mechanisms and using different types of data.

2.1.3 HAR using Depth Data

The depth modality can be used for HAR because it reliably captures the 3D structural and geometric shape of human subjects despite variations in colour and texture (Sun et al., 2023). Converting 3D data into a 2D image is the crux of depth map construction. In the existing literature, there are several works on HAR based on depth data of the scene. The majority of these techniques relied on depth maps generated by inexpensive sensors such as Kinect (Li et al., 2010; Yang and Tian, 2012; Oreifej and Liu, 2013; Yang and Tian, 2014). The majority of existing depth-based action recognition techniques rely on global features such as space-time volume and silhouette data (Wang et al., 2020). The existing depth-based approaches can be broadly classified as either DL-based or handcrafted feature-based methods (Yang and Tian, 2014).

Various handcrafted feature-based methods have been proposed in the literature. (Oreifej and Liu, 2013) introduced the use of a histogram representing the distribution of the surface normal orientation in the space of time, depth, and spatial coordinates for representing the action in depth data. (Yang and Tian, 2014) proposed clustering hyper-surface normals in a depth sequence to create the poly-normal, which is then used to jointly characterize the local motion and shape features required for HAR. (Xia and Aggarwal, 2013) introduced a filtering technique for extracting STIPs from depth videos that effectively suppress noisy measurements. Further, a depth cuboid similarity feature is used to characterize the local 3D depth cuboid surrounding the depth STIPs with an adjustable supporting size. (Rahmani et al., 2016) proposed the extraction of the histogram of oriented principal components descriptor by directly processing the point-clouds to solve issues arising from noise, viewpoint, and action speed variations.

DL based methods demonstrated better performance in HAR based on depth data. There are several approaches on HAR using DL and depth data (Rahmani and Mian, 2016; Shi and Kim, 2017; Zhang et al., 2018), proposed for HAR using DL and depth data. Due to the success of handcrafted Depth Motion Maps (DMM), a weighted hierarchical DMM in a DL framework is proposed by (Wang et al., 2016). Using segmented sequence of depth maps, (Wang et al., 2018) developed three effective representations of depth sequences, namely: dynamic depth images, dynamic depth normal images, and dynamic depth motion normal images for both isolated and continuous action recognition. These descriptors are learnt by a CNN-based DL model for accurate HAR. To automatically encode spatio-temporal patterns from depth sequences without pre-processing a 3D CNN model is proposed by (Sanchez-Caballero et al., 2022). A ConvLSTM -based DL model for learning spatio-temporal features from sequence of raw depth maps is introduced in (Sanchez-Caballero et al., 2020). The summary of key depth-based HAR works is described in Table 2.3.

Table 2.3: Summary of key existing HAR works based on depth data.

Authors	Methodology	Remarks
(Yang et al., 2012)	Project depth maps onto three orthogonal planes and accumulate global activities across video sequences to create the DMM. Histograms of Oriented Gradients (HOG) are computed from DMM to represent action.	The compact and discriminative representation captures global activities from front/side/top views.
(Song et al., 2014)	Proposed local depth map feature describing action's spatio-temporal details.	A feature that is approximately object-centered, making it more tolerant of variations. Robust to view variations.

Table 2.3 Continued from previous page.

Authors	Methodology	Remarks
(Liang et al., 2016)	Extracts a multilayered depth motion feature and applies a multi-scale HOG descriptor to capture the local temporal change of human motion and action spatial structure	Effective for small HAR sample sizes. This method makes the sparse coding coefficients sufficiently discriminatory for classifying actions that are similar.
(Yang and Tian, 2017)	To jointly characterise local motion and shape information, surface normals are extended to poly-normal by assembling local neighbouring hyper surface normals from a depth sequence. Adaptive spatio-temporal pyramid subdivides a depth video into space-time cells to capture spatio-temporal features.	The framework can be easily adapted for use in any depth sequence that is aligned with a joint trajectory. When there are large variations in both space and scale, this approach works well.
(Ahmad et al., 2019)	Raw depth maps are processed using CNN model.	As entire set of depth maps are used, thus needs more computations.
(Weiyao et al., 2019)	A Multilevel Frame Select Sampling (MFSS) method is proposed to generate three levels of temporal samples from the input depth sequences. The motion and static mapping, method is applied to generate the representation of MFSS sequences.	Three temporal levels can achieve better recognition $Accuracy$ when compared with other temporal levels. Need to test with large-scale data set.

2.1.4 Multi-modal HAR

The availability of visual data in multiple modalities has led to the creation of HAR systems based on the fusion of data modalities. (Fan et al., 2020) proposed a method that combines RGB and skeleton data, in which context-aware cross-attention module to extract joints that are closely relevant to the context information and are more insightful. The context information branch contains two branches that are applied to RGB data. (Gu et al., 2020) proposed a method for HAR that utilizes both low-level characteristics and high-level contextual information. Moreover, this utilizes data in three modalities: skeleton, depth, and RGB, to extract the required data for HAR. (Kamel et al., 2019) suggested a deep CNN-based model with score fusion operations using depth and skeleton data. Two action descriptors were introduced as images to support HAR from the front view. (Romaissa et al., 2021) proposed fusion of RGB, depth, and skeleton data, further the LSTM model is used to learn the spatio-temporal features for HAR. The summary of most significant multi-modal works on HAR is reported in Table 2.4.

Table 2.4: Summary of key existing HAR works based on multiple data modalities.

Authors	**Modality**	**Methodology**	**Remarks**
(Kamel et al., 2019)	Skeleton, Depth	The action descriptors based on skeleton and depth data are trained with 3 channel CNN, followed by fusion operations.	Scope for improving the performance by focusing on more important features.
(Yang et al., 2020)	Skeleton, Depth	Attempt to reduce the redundancy in depth maps and captures spatial motion states. Generated action descriptor using skeleton data.	The method establishes an inherent relationship between the labels.
(Fan et al., 2020)	RGB, Skeleton	Used attention mechanism to select informative joints. It is combined with context information extracted from RGB data.	Using entire RGB data for context information needs to process huge data.
(Singh et al., 2020)	Depth, RGB	A depth and RGB sequence constituting the action video is used to generate dynamic images. Using a multi-stream CNN model with score fusion, these dynamic images are learned.	Performance can be improved by focusing on to the most important period during an action.
(Romaissa et al., 2021)	Skeleton, Depth, RGB	Using a sequence of RGB, depth, and skeleton frames, separate dynamic images are constructed to represent the action. These dynamic images are trained with a DL model comprised of CNN, and LSTM.	Utilizing three distinct data modalities for action recognition is computationally intensive.

Table 2.4 Continued from previous page.

Authors	**Modality**	**Methodology**	**Remarks**
(Cheng et al., 2022)	Depth, RGB	A powerful 2D-CONVNet model with a cross-modality compensation module to discover complementary discriminative features from two modalities is introduced.	Feature fusion improved the performance, however RGB data demands more resources.

2.1.5 Datasets for HAR

Several datasets are made publicly available for research on HAR. This thesis work mainly focuses on RGB, skeleton, and depth modality datasets for HAR. As, one of the task is domain specific we used custom RGB dataset. Rest of the works are carried using publicly available datasets captured using Kinect depth sensor. The statistics of some of the benchmark datasets for HAR are listed in Table 2.5.

Table 2.5: Datasets for Human Action Recognition.

Dataset	**Authors**	**Data Modality**	**Samples**	**Classes**	**Subjects**	**Views**
MSRAction-3D	(Li et al., 2010)	Depth, Skeleton	567	20	10	1
UT-Kinect	(Xia et al., 2012)	RGB, Depth Skeleton	200	10	10	4
SBU-Kinect Interaction	(Yun et al., 2012)	RGB, Depth, Skeleton	300	8	7	1
MSRDaily Activity3D	(Wang et al., 2012)	RGB, Depth, Skeleton	320	16	10	1
N-UCLA	(Wang et al., 2014)	RGB, Depth, Skeleton	1494	10	10	Variety (3 cameras)

Table 2.5 Continued from previous page.

Dataset	**Authors**	**Data Modality**	**Samples**	**Classes**	**Subjects**	**Views**
UTD-MHAD	(Chen et al., 2015)	RGB, depth, skeleton	861	27	8	1
NTU-RGB+D	(Shahroudy et al., 2016)	RGB, Depth, Skeleton, IR	56880	60	40	80 (3 cameras)
NTU-RGB+D120	(Liu et al., 2019)	RGB, Depth, Skeleton, IR	114480	120	106	155 (3 cameras)

This work is conducted on two single-view datasets, namely: MSRAction3D and UTD-MHAD, and one multi-view dataset: NTU RGB+D using skeleton and depth data. Tables 2.6, 2.7, and 2.8 gives the action labels in MSRAction3D, UTD-MHAD, and NTU RGB+D, respectively.

Table 2.6: Actions in MSRAction3D.

Sl. No.	**Label**	**Sl. No.**	**Label**	**Sl. No.**	**Label**	**Sl. No.**	**Label**
1	High arm wave	2	Horizontal arm wave	3	Hammer	4	Hand catch
5	Forward punch	6	High throw	7	Draw cross	8	Draw tick
9	Draw circle	10	Hand clap	11	Two-hand wave	12	Side boxing
13	Bend	14	Forward kick	15	Side kick	16	Jogging
17	Tennis swing	18	Tennis serve	19	Golf swing	20	Pic-up and throw

Table 2.7: Actions in UTD-MHAD.

Sl. No.	**Label**	**Sl. No.**	**Label**	**Sl. No.**	**Label**
1	Right arm swipe to the left	2	Right arm swipe to the right	3	Right hand wave
4	Two hand front clap	5	Right arm throw	6	cross arms in the chest
7	Basketball shoot	8	Right hand draw x	9	right hand draw circle-clockwise
10	Right hand draw circle-counter clockwise	11	Draw triangle	12	Bowling right hand
13	Front boxing	14	Baseball swing from right	15	Tennis right hand forehand swing
16	Arm curl two arms	17	Tennis serve	18	Two hand push
19	Right hand knock on door	20	Right hand catch an object	21	Right hand pick up and throw
22	Jogging in place	23	Walking in place	24	Sit to stand
25	Stand to sit	26	Forward lunge	27	Squat

2.2 Human Identification

Automatically identifying a person from a group of people is one of the most important tasks for ensuring a comfortable and safe life in the era of smart environments. Vision-based gait recognition identifies a person by analysing camera-collected visual data of a pedestrian's walking pattern. Several methods, such as (Zhang et al., 2019; Vrigkas et al., 2015), can be found in the literature for Human Identification; however, each of these methods relies on human cooperation for identification. In unobtrusive human identification systems, the features are collected from the human without his/her knowledge. Vision-based gait features can be collected from far without the subject's awareness (Khamsemanan et al., 2018; Singh et al., 2018; Sepas-Moghaddam and Etemad, 2023). So HI plays an important role in smart surveillance systems.

There are two kinds of conventional gait identification algorithms, namely: model-free (Han and Bhanu, 2006; Huang and Boulgouris, 2012) and model-based (Tafazzoli and Safabakhsh, 2010; Li et al., 2020; Liao et al., 2022; Zheng et al., 2022). Model-free techniques are often known as appearance-based techniques. In contrast, model-based methods aim to reconstruct a person's three-dimensional model. Here, gait data

Table 2.8: Actions in NTU RGB+D.

Sl. No.	Label	Sl. No.	Label	Sl. No.	Label
1	Drink water	2	Eat meal/snack	3	Brushing teeth
4	Brushing hair	5	Drop	6	Pickup
7	Throw	8	Sitting down	9	Standing up
10	Clapping	11	Reading	12	Writing
13	Tear up paper	14	Wear jacket	15	Take off jacket
16	Wear a shoe	17	Take off a shoe	18	Wear on glasses
19	Take off glasses	20	Put on a hat/cap	21	Take off a hat/cap
22	Cheer up	23	Hand waving	24	Kicking something
25	Reach into pocket	26	Hopping	27	Jump up
28	Make a phone call/answer phone	29	Playing with phone/tablet	30	Typing on a keyboard
31	Pointing to something with finger	32	Taking a selfie	33	Check time
34	Rub two hands together	35	Nod head/bow	36	Shake head
37	Wipe face	38	Salute	39	Put the palms together
40	Cross hands in front	41	Sneeze/cough	42	Staggering
43	Falling	44	Touch head	45	Touch chest
46	Touch back	47	Touch neck	48	Nausea or vomiting condition
49	Use a fan /feeling warm	50	Punching/slapping other person	51	Kicking other person
52	Pushing other person	53	Pat on back of other person	54	Point finger at the other person
55	Hugging other person	56	Giving something to other person	57	Touch other person's pocket
58	Handshaking	59	Walking towards each other	60	Walking apart from each other

is streamlined into a known structure such as skeletons or body structures before the extraction of features. Prior to the development of sensor devices, model-based approaches had not been used extensively (Khamsemanan et al., 2018). This situation has changed as a result of new technologies, specifically the development of Microsoft Kinect and its SDK (Zhang, 2012). This thesis primarily concentrates on HI based on gait data captured using depth sensors. Below we discuss some of approaches using 2D silhouette images.

2.2.1 2D Gait-based Human Identification

Here, the human identification is made based on the features constructed from the image sequence of human walking. In (Yoo et al., 2008) used a model-based approach wherein the human body points from the gait are used to model the human body as 2D stick figures. A number of existing approaches used image silhouettes for feature extraction such as (Li et al., 2008; Liang Wang et al., 2004; Wan et al., 2018). Gait Energy Image (GEI) (Han and Bhanu, 2006) is a model-free gait recognition method that employs the average silhouette image as gait features. It is regarded as a standard algorithm for model-free gait recognition and is one of the most popular due to its simplicity and efficacy. The GEI used by (Ma et al., 2017) for feature extraction and followed by neural network based classification. Where as (Ju Han and Bir Bhanu, 2006) used similarity measurement of GEI for human identification. In (Li et al., 2008), human silhouette image is divided into seven parts and mainly studied the contribution of each part for gait recognition. (Liang Wang et al., 2004) segmented human body into fourteen parts and used joint angle trajectories for human identification. (Tang et al., 2017) suggested gait recognition using 3D parametric body models are morphed by pose and shape deformation from a template model using 2D gait silhouette sequence.

2.2.2 3D Gait-based Human Identification

Among 2D model-based gait recognition, most of the methods established skeleton model using image sequence but it is affected by illumination, clothing, etc. (Wang et al., 2016). So, in recent years with the development of 3D cameras, 3D gait recognition also started gaining importance as they have better performance in view variance and further, the data need to be processed is less compared to 2D. 3D skeleton joint data provided by Kinect depth sensor eliminates the need for complex procedures of building a model from visual data streams (Deng and Wang, 2019).

Several works (Choi et al., 2019; Khamsemanan et al., 2018; Deng and Wang, 2019; Bari and Gavrilova, 2019; Limcharoen et al., 2020) have been proposed for 3D skeleton-based gait recognition by exploiting various static and dynamic gait-specific features. Some techniques extracted features from raw skeleton data using direct deep learning models. In contrast, some methods employ a separate step for feature extraction. In addition, these features are classified using machine learning or deep learning models. Below are examples of works that fall under both of these categories. (Yang et al., 2016) generated a set of features using relative joint distances. Then, the distance features and anthropometric features are combined to create the final feature vector. Additional K-Nearest Neighbors (KNN) and majority voting are used for HI. A deep neural network with joint relative cosine similarities and triangle areas based 3D gait recognition is proposed in (Bari and Gavrilova, 2019). (Sun et al., 2018) created a feature set comprising of the length of specific skeletons and swing angles of limbs and further used Nearest Neighbor (NN) classification. (Choi et al., 2019) performed person identification by frame-level discriminative scores, but the time required to identify the person increases with number of frames. (Li et al., 2017) proposed a LSTM model to learn raw skeleton data from the sequence for person's identification. However, there is a possibility that performance will suffer if the skeleton is noisy. (Hosni and Amor, 2020) proposed a geometric deep CNN encoding-decoding framework for 3D gait recognition. LSTM model is proposed to learn skeleton information from each frame. (Li et al., 2017). This method used raw data without accounting for gait-specific features. The summary of key works on HI using 3D data inspired this thesis work is given in Table 2.9.

Table 2.9: Summary of key existing HI works based on data from Kinect depth sensors.

Authors	**Modality**	**Methodology**	**Remarks**
(Haque et al., 2016)	Depth	Proposed a recurrent attention model to identify the important spatio-temporal regions for the person identification problem from depth video of walking sequence.	Use of attention gives importance to discriminative features.

Table 2.9 Continued from previous page.

Authors	**Modality**	**Methodology**	**Remarks**
(Karianakis et al., 2017)	Depth	Proposed reinforced temporal attention on frame-level features to capture the temporal information from video sequences for person re-identification from depth data.	The reinforced Temporal Attention unit is independent of the network architecture and focuses on discriminative features for person identification.
(Wu et al., 2017)	Depth, Skeleton	Exploited depth information for invariant body shape and skeleton information regardless of illumination and color change. Also, used a kernelized implicit feature transfer scheme to estimate the Eigen depth from RGB image.	Used a kernelized implicit feature transfer scheme to estimate the Eigen depth from RGB image in the absence of depth image.
(Khamsemanan et al., 2018)	Skeleton	The posture-based features of each frame are classified using Machine Learning (ML) techniques, and the probability score of each frame is combined to make a final decision.	An attempt to handle varying viewpoint-related issues. The classification results of noisy frames affect the performance.
(Bari and Gavrilova, 2019)	Skeleton	Two view-invariant geometric features, joint relative cosine dissimilarity and triangle area are extracted from the frames in the gait cycle and trained using a neural network model.	Complex deep learning model, which has more trainable parameters thus need more computation.
(Liu et al., 2019)	Skeleton	Two separate deep learning models with LSTM and CNN are proposed to capture spatial and temporal information by processing skeleton gait energy image and joint angles.	Can improve the performance by focusing on to most discriminative features.

Table 2.9 Continued from previous page.

Authors	Modality	Methodology	Remarks
(Limcharoen et al., 2020)	Skeleton	Focused on multi-viewpoint challenges in gait recognition. Feature vectors contain the joint replacement coordinates using a set of selected frames and a CNN-based DL model is proposed.	Few local joint movements were considered. To improve the performance, there is a need to consider the entire body movement data.
(Huynh-The et al., 2020)	Skeleton	Extracts geometric distance and orientation features. A CNN model is proposed for feature learning and classification.	Features are accumulated over multiple frames to compute gait sequence statistics.
(Xu et al., 2021)	Skeleton	Proposed Local Graphical Skeleton Descriptor (LGSD) for extracting the geometrical patterns of the skeleton sequence. These are processed using Dual-stream CNN-based DL model for classification.	Used only the local descriptors of the skeleton. It can be improved further by learning the end-to-end features.
(Limcharoen et al., 2021)	Skeleton	It targets the rhythm of movements in 22 different regions of the human body using region-specific LSTM models. Outputs from these 22 LSTM models are combined to learn the relations among regions.	Captures the relation among different regions effectively. Increased computational complexity as 22 different models are used.

2.2.3 3D Gait Datasets for Human Identification

Several datasets are made publicly available for research on HI. This thesis work focuses on skeleton data based gait sequences. The details about some of the benchmark datasets used in this work for HI are listed in Table 2.10.

Table 2.10: Summary of skeleton-based benchmark gait datasets.

Dataset	UPCV1	UPCV2	KGBD	KS20	IAS-lab
Authors	(Kastaniotis et al., 2015)	(Kastaniotis et al., 2016)	(Andersson and Araujo, 2015)	(Nambiar et al., 2017a)	(Munaro et al., 2014)
Subjects	30	30	164	20	11
Samples	150	300	822	300	11+11+11
Multi-View	No	No	No	Yes	No
Walking Direction	straight line	straight line	semi circular	straight	arbitrary

2.3 Outcome of Literature Review

Existing research has established the importance of smart surveillance systems in smart environments such as smart cities, smart campuses, etc. In addition, HAR and human identification systems are two essential functions of intelligent surveillance systems in providing a secure and comfortable living environment. Numerous studies have been conducted on HAR and HI to address various problems. Based on the extensive literature review, the following research gaps are identified.

Research Gaps

- Every day, surveillance cameras generate massive amounts of video data. Consequently, analyzing and storing data for future use is extraordinarily challenging. Vision-based surveillance systems have no efficient way of storing video data based on interpreting human actions in the scene.

- Most existing works have focused on recognizing single human actions from images depicting one human action. However, in reality, an image may contain multiple actions. Consequently, precise localization and recognition of multiple human actions in a scene are challenging.

- There is no widely adopted architecture for localization and recognition of multiple human actions in images and videos.

- Most of the human actions are application domain-specific. Consequently, it is crucial for surveillance system applications to utilize domain-specific human action data. For instance, no established datasets for student actions on campus is found in the literature.

- There is diversity in the way different people perform the same action, for instance, in terms of speed, duration, and so on. The subject (person) independent HAR needs to be improved for effective people monitoring.

- Human actions appear differently from different viewpoints, and there is no robust method that can handle variation in viewpoints for HAR.

- Vision-based surveillance systems require unobtrusive identification of humans. Human gait is biometric that can be used for unobtrusive identification. The gait of a person appears differently from different viewpoints. There is a need for effective methods to improve gait-based human identification.

- The works on a smart surveillance system in the smart environment are still in the infancy stage.

- Varying light conditions and occlusions can affect both human actions as well as human appearance features, and they are not explored much in the existing works related to HAR and Human identification.

- Vision-based surveillance systems require unobtrusive identification of humans. Human gait is biometric that can be used for unobtrusive identification. The gait of a person appears differently from different viewpoints. There is a need for effective methods to improve gait-based human identification.

- The recognition of complex human actions from video and identifying humans from their gait requires more resources than approaches based on still/single images. There are no methods for effectively reducing the number of features and, in turn, the demand for resources without sacrificing performance.

- Various modalities of video-based gait and human action data have recently been generated. The works that exploit the benefits of combining data modalities for HAR and HI are yet to be extensively investigated.

2.4 Problem Statement

An unobtrusive automated analysis of videos for interpreting the content and identifying the people in the scene is a crucial requirement for providing a safe, high-quality, and comfortable living in smart environments. Based on the literature review in this direction and the research gaps identified, the research problem is stated as follows:

"Design and develop a vision-based unobtrusive context-aware Human Identification and Action Recognition system for smart environments using deep learning techniques."

2.5 Research Objectives

The following four research goals are addressed in this thesis based on the identified research gaps and problem statement:

1. To design and develop a context-aware, view-invariant Human Action Recognition system based on RGB data captured spontaneously.

2. To design and develop an action representation and classification model for recognizing human actions based on single-view skeleton data.

3. To design and develop an efficient feature extraction and classification system for human identification using skeleton-based gait data from single-/multi-view scenarios and novel deep learning models.

4. To design and develop an effective representation of spatio-temporal features of 'human action' & 'human gait cycle', and classification model for single-/multi-view 'human action recognition' & 'human identification' using skeleton & depth data fusion and deep learning models.

2.6 Summary

This chapter discussed current state-of-the-art techniques for HAR and gait-based HI. The various uni-modal and multi-modal approaches for HAR based on visual data are presented in detail. In addition, numerous HI approaches employing gait data for unobtrusive identification of a person with an emphasis on visual data in skeleton format are discussed. Further, information regarding the benchmark datasets for HAR and HI is provided.

Based on the literature review findings, the challenges in the fields of HAR and HI are articulated. In addition, the problem statement and research objectives based on the outcome of literature review are presented in detail. In the following chapters, HAR and HI-specific solutions are provided for the challenging issues raised in this chapter. The next chapter describes a domain-specific HAR system for smart computer laboratories.

Chapter 3

Context-aware Human Action Recognition System for Computer Laboratories of Smart Campus

People engage in a variety of actions in their daily lives. Most of these actions have strong ties to the context or environment in which they are carried out. This thesis work aims to recognize the human actions performed in a computer laboratory of a smart campus environment using the RGB image data.

3.1 Smart Campus

Smart campus is defined as an integrated system with cooperation and self-adjustment capabilities, based on the Internet of Things (IoT), that enables a wise, intelligent teaching, learning, and living environment suitable for teaching, scientific research, and management, among other applications (Du et al., 2016). The specifications and implementation of a smart campus are tailored to each institution's particular needs. The advancements in IoT technology, DL, and ML, a smart campus allows campus staff to concentrate on their primary responsibilities while automating as many processes as possible to aid in decision-making.

Several methods utilizing image and video analysis have been proposed to decode the student's mood, level of interest, and concentration on learning. (Candra Kirana et al., 2018), for example, proposed a method for facial emotion recognition in a learning environment based on the Viola-Jones algorithm. A facial expression database is being compiled in online learning environments to meet the needs of automatic academic emotion inference (Bian et al., 2019). (Whitehill et al., 2014) proposed facial expression-based methods for automatically detecting the student's interest. (Bosch and D'Mello, 2019) developed a video-based, facial-features-based mind-wandering detector for classroom and laboratory use.

There has been a dramatic increase in the number of people enrolling in college courses in recent years. Students' activities must be monitored to create a more conducive campus teaching and learning environment. In the conventional method of campus surveillance, human resources are typically utilized. In the past few years, campuses typically installed a large number of surveillance cameras in both indoor and outdoor environments for a variety of reasons. The campus can use the footage from these cameras for safety, management, and planning. These cameras can produce vast and intricate amounts of video data. There is a need for system that automates the interpretation of human actions in these videos.

Contributions:

To ensure the most efficient use of resources in the computer lab, we require a monitoring system for student activities within the lab. This work mainly focuses on locating and recognizing the students' actions inside the computer lab environment in Indian context. The primary contributions of this work are as listed below.

- Created a dataset containing five distinct student actions for multiple students within a single image frame using the video captured by Closed Circuit Television (CCTV) cameras in computer science laboratories.
- Recognizing and localizing multiple human actions in a single RGB image frame using a DL model based on a transfer learning approach.
- Implemented a frame reduction strategy for analyzing video captured by CCTV cameras installed in computer labs to monitor students as efficiently and precisely as possible.

3.2 Proposed Methodology

The architecture of the proposed method for the localization and recognition of multiple student actions in computer laboratories using RGB image frames is depicted in Figure 3.1. It primarily consists of the following three subsystems:

- Dataset Preparation
- Training Deep Learning Model
- Action Recognition

The 'Dataset Preparation' subsystem creates a new dataset STUDENT_ACTION for the proposed work. In the next subsystem, transfer learning is applied to the pre-trained YOLOv3 (Redmon and Farhadi, 2018) model for the proposed work. Finally, the newly trained model is used in the 'Action Recognition' subsystem to aid in student's action recognition. Since the content of consecutive image frames from a video are similar, we also proposed a method for reducing the total number of image frames to be processed in a video captured by CCTV camera. A detailed explanation of each of these subsystems is given in the following subsections.

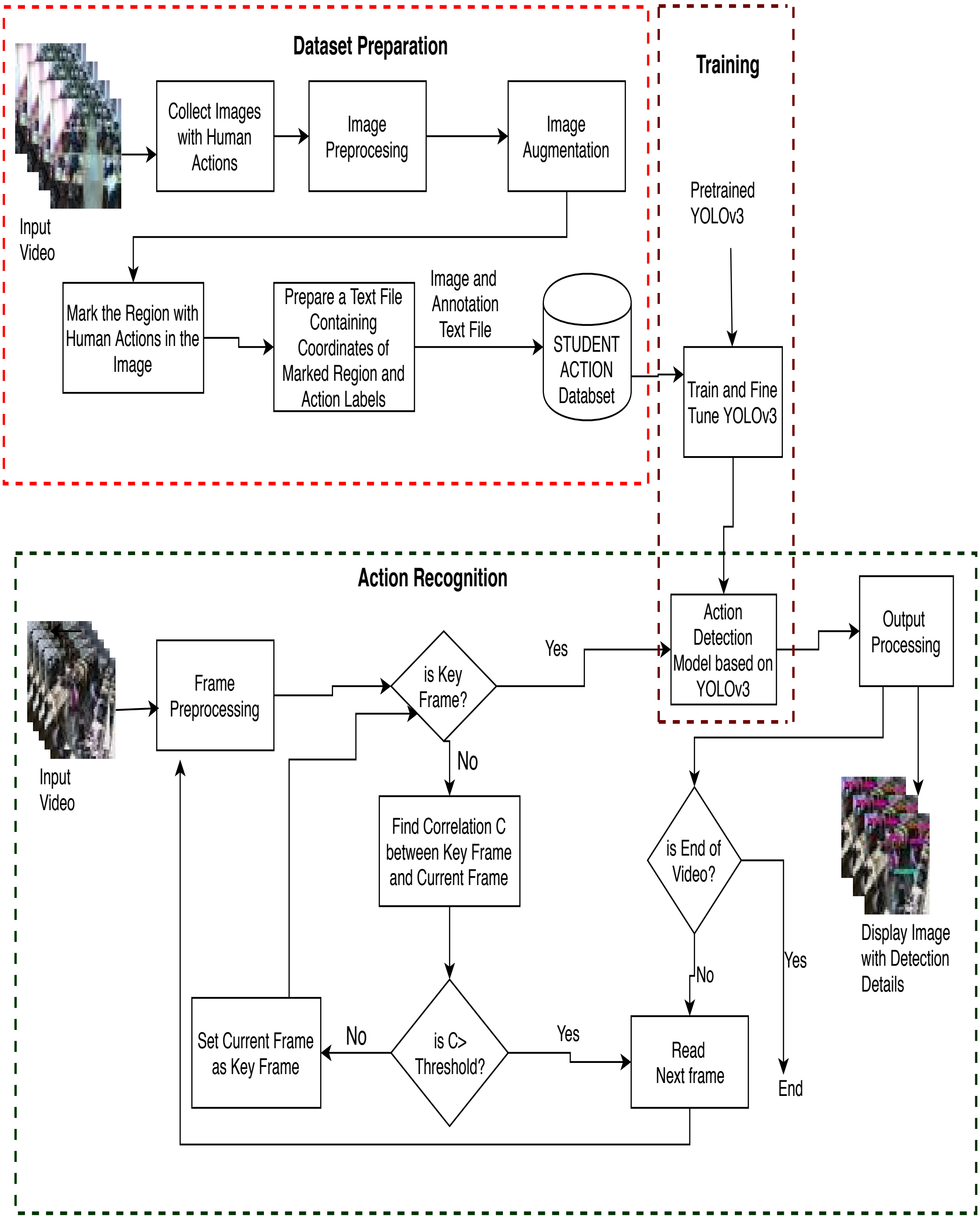

Dataset Preparation
Training
Input Video
Collect Images with Human Actions
Image Preprocesing
Image Augmentation
Pretrained YOLOv3
Mark the Region with Human Actions in the Image
Prepare a Text File Containing Coordinates of Marked Region and Action Labels
Image and Annotation Text File
STUDENT ACTION Databset
Train and Fine Tune YOLOv3
Action Recognition
Input Video
Frame Preprocessing
is Key Frame?
Yes
No
Action Detection Model based on YOLOv3
Output Processing
Find Correlation C between Key Frame and Current Frame
is End of Video?
Yes
No
Display Image with Detection Details
Set Current Frame as Key Frame
No
is C> Threshold?
Yes
Read Next frame
End

Figure 3.1: Proposed architecture for student's action recognition using RGB data.

3.2.1 Dataset Preparation

To ensure that the proposed system can effectively track student's actions inside the smart campus computer laboratory, we initially need to define a dataset that adequately characterizes these actions. Below is a detailed description of the steps involved in creating the STUDENT_ACTION dataset.

3.2.1.1 Collect Images with Human Actions

Image frames are extracted from videos captured by CCTV cameras. To identify human actions from static images by analyzing factors such as gestures, posture, head position, and objects near the human, we focused on only those images which contain human actions for preparing the dataset.

3.2.1.2 Image Pre-processing

Because students' actions are the primary focus of the proposed system, the portion of the image is cropped to eliminate the parts along the image's four sides that do not contribute to student's activity. Then, each image is scaled to precisely 416×416 pixels.

3.2.1.3 Image Augmentation

Table 3.1: Data augmentation techniques applied in the dataset.

Augmentation Technique	Description
Gaussian Blur	It is a type of image-blurring filter that uses a Gaussian function for calculating the transformation to apply to each pixel in the image.
Median Blur	The central element of the image is replaced by the median of all the pixels in the kernel area.
Bilateral Filter	Intensity of each pixelis replaced with a weighted average of intensity values from nearby pixels.
Dilation	Morphological image transformation.
Changing Contrast	Image with modified contrast value.

The image data augmentation techniques listed in Table 3.1 are used to increase the number of images in the dataset to improve the performance of the DL model in localization and recognition of student's actions.

3.2.1.4 Data Annotation

Since the primary purpose of the proposed system is the localization and recognition of student's action, the data annotation consists of two primary tasks: labeling the human actions and determining the bounding box coordinates for each action in the image. In this instance, we must be able to detect actions, identify the type of action being carried out, and pinpoint the exact location within the image where the action is occurring. A single image can contain multiple actions in different parts of the image. Therefore, all such actions in images are identified and annotated manually. The method described in (Gu, 2019) is used to generate action labels and bounding boxes corresponding to the identified actions. The action label and the coordinates of associated bounding boxes are recorded in the annotation text file for each frame. The dataset stores the text file containing annotations and the corresponding image frame. The following is an illustration of the format used for annotations in the proposed method:

One row for one image;

Row format: image-file-path box1 box2 ... boxN

Box format: xmin,ymin,xmax,ymax,action_label

3.2.2 Training Deep Learning Model

The YOLOv3 (Redmon and Farhadi, 2018) model has been configured for the proposed work. Using the pre-trained weights of YOLOv3, and the newly created dataset, the model for the proposed action recognition task is trained and fine-tuned. The proposed method employs the newly generated weights for the 'Action Recognition' subsystem. During training, the K-fold cross-validation method is employed to test each sample in the dataset.

3.2.2.1 Overview of YOLOv3

YOLOv3 is one of the most rapid and effective object detectors, with excellent real-time performance. YOLOv3 initially represents an improvement over YOLO (Redmon et al., 2016). YOLO partitions the image input into an $S \times S$ grid, where, each grid cell associated with a single object. It can predict a fixed number of bounding boxes, with a box confidence score associated with each bounding box. Therefore, the information to be predicted for each bounding box consists of 5 values: *(x, y, w, h, box confidence*

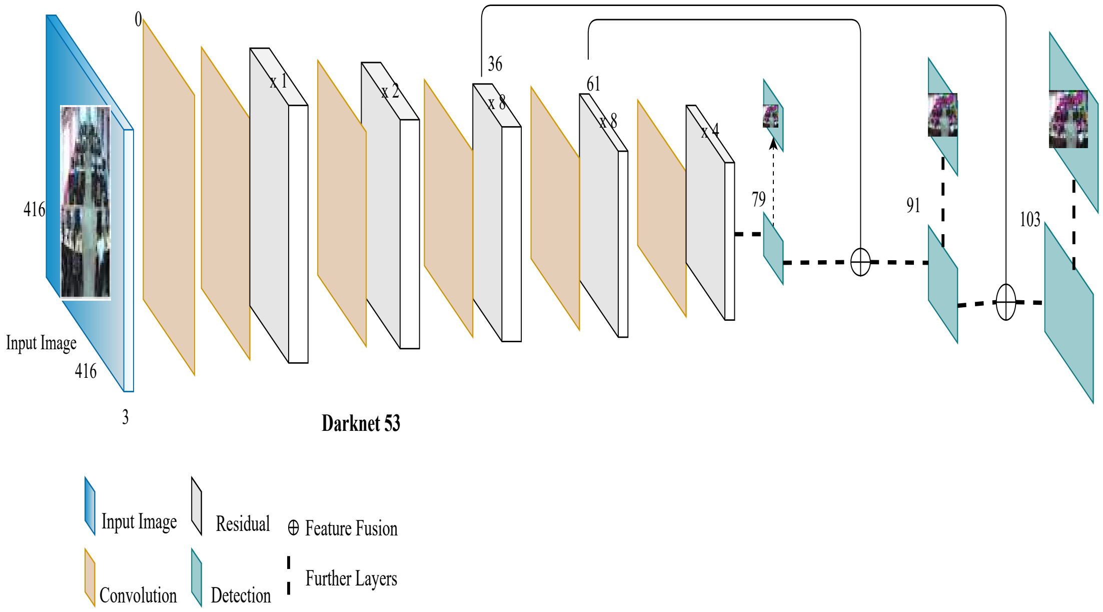


Figure 3.2: YOLOv3 feature extractor (Darknet-53) (Redmon and Farhadi, 2018)

score), where, the *(x, y)* coordinates represent the centre of the box relative to the grid cell location and *(w, h)* are the box dimensions, and the box confidence score indicates the likelihood that the box contains an object and the precision of the bounding box. The final step involves in calculating the class confidence score for each prediction box as the product of the box confidence score and the conditional class probability.

Figure 3.2 illustrates the architecture of the YOLOv3 feature extractor. The model employs 3 and 1×1 convolutional layers in succession. YOLOv3 employs various layers, including convolutional layers, shortcut layers, route layers, and the YOLO detection layer. Where the shortcut layer provides the skip connection, the up-sampling layer increases the resolution of the previous layer's feature maps. Route layer has a layer attribute with one or two values; if it has one value, outputs feature maps of the layer indexed by the value; otherwise, concatenated feature maps of layers indexed by its values. Finally, the detection layer of YOLOv3 specifies the anchors (bounding boxes by default) used during detection.

3.2.3 Action Recognition

This subsystem is responsible for analyzing the CCTV footage captured by computer labs' surveillance cameras and identifying and locating the student's actions. Since the

YOLO model prioritizes both speed and efficiency, it is well-suited for real-time applications. The trained YOLOv3-based model has the ability to recognize and localize student's actions within a single image frame. As successive image frames in a video are highly similar, in the proposed work to accelerate video analysis, the number of frames to be processed is reduced by employing a technique called template matching.

Algorithm 3.1: Algorithm for Action Recognition Subsystem

```
Input: Video Stream from CCTV camera
Output: Image frames with action labels and bounding boxes around detected
        student's actions
Initialize: flag = True
  Extract first image frame from video stream
  Pre-process the image frame and set it as keyframe and current_frame

while flag do
    if keyframe and current_frame are same then
        Detect actions in current_frame using fine tuned action detection model
          Display image along with detection labels and bounding boxes around
          recognized actions
          if end of video stream then
            flag = False
    else
        find Template Match between keyframe and current_frame and set to C
          if $C \leq threshold$ then
            keyframe = current_frame
        else
            skip current_frame
              Extract and pre-process next frame and set it as current_frame
        end
    end
end
```

Initially, this computes the template match between two frames. If the template match exceeds the threshold, the frame is skipped, and the next frame is read. The template matching threshold in the proposed system is set to 0.997%. Experiments reveal that the time required for template matching is less than the time required to process the frame for action detection. Template matching consumed an average of 0.099 seconds in the system where experiments are conducted, whereas action detection requires 0.90 seconds. Additionally, it is observed that actions from skipped frames are present in non-skipped frames. This drastically reduces the number of frames to be processed for video analysis. Algorithm 3.1 describes the steps involved in action recognition. Algorithm 3.1 can be executed in $O(n)$ time, where n is the number of video image frames. Template matching is determined using the Equation (3.1) in Algorithm 3.1 according to the method described in (Vision, 2019).

$$R(x,y) = \frac{\sum_{x',y'}(T(x',y') \cdot I(x+x',y+y'))}{\sqrt{\sum_{x',y'} T(x',y')^2 \cdot \sum_{x',y'} I(x+x',y+y')^2}}, \tag{3.1}$$

Where I, T, and R represent the image, template, and result, respectively. The parameters (x, y) denote the pixel positions. x' ranges from zero to the width of the template, and y' ranges from zero to the height of the template. In the proposed algorithm, I and T represent the current keyframe and resents the current frame extracted from the video, respectively.

3.3 Details about Dataset

As the proposed system is intended to monitor student's actions in the computer lab of a smart campus, initially a data set that best describes student's actions in the specified domain is constructed. The STUDENT_ACTION dataset was created using image frames extracted from videos captured (during 2017 and 2018) by CCTV cameras installed in computer laboratories at the Department of Information Technology, National Institute of Technology Karnataka. The STUDENT_ACTION dataset includes the action labels associated with Indian engineering college laboratory standards. The dataset consists of five distinct action labels, which are all described in Table 3.2. Initially, the dataset was constructed using 688 original image frames captured by different CCTV cameras. To enhance the recognition performance of the proposed methodology, the model is trained using images of varying quality. The number of images in the dataset has been increased to 6,500 through the use of various augmentation techniques. Figures 3.3, and 3.4 show the sample original and augmented images in the dataset.

Each frame captured by the CCTV cameras installed in computer labs depicts various human actions dispersed throughout the image. The number of samples for action label "*Engaged*" is extremely high because the vast majority of students in the laboratory will be engaged in activities such as using computers or reading books. The

Table 3.2: Actions in RGB image-based STUDENT_ACTION dataset.

Action Label	**Definition**
Discussion	Two or more persons sitting or standing together and facing to each other.
Engaged	A person looking at book or person looking at monitor or looking at keyboard and hands in keyboard.
Sleeping	A person bent towards the computer table and kept head on the table.
Eating	A person holding something in hand close to mouth or holding bottle in hand or near the mouth.
Using_Smart_Phone	A person looking at smartphone in hand.

Table 3.3: Distribution of actions among the frames.

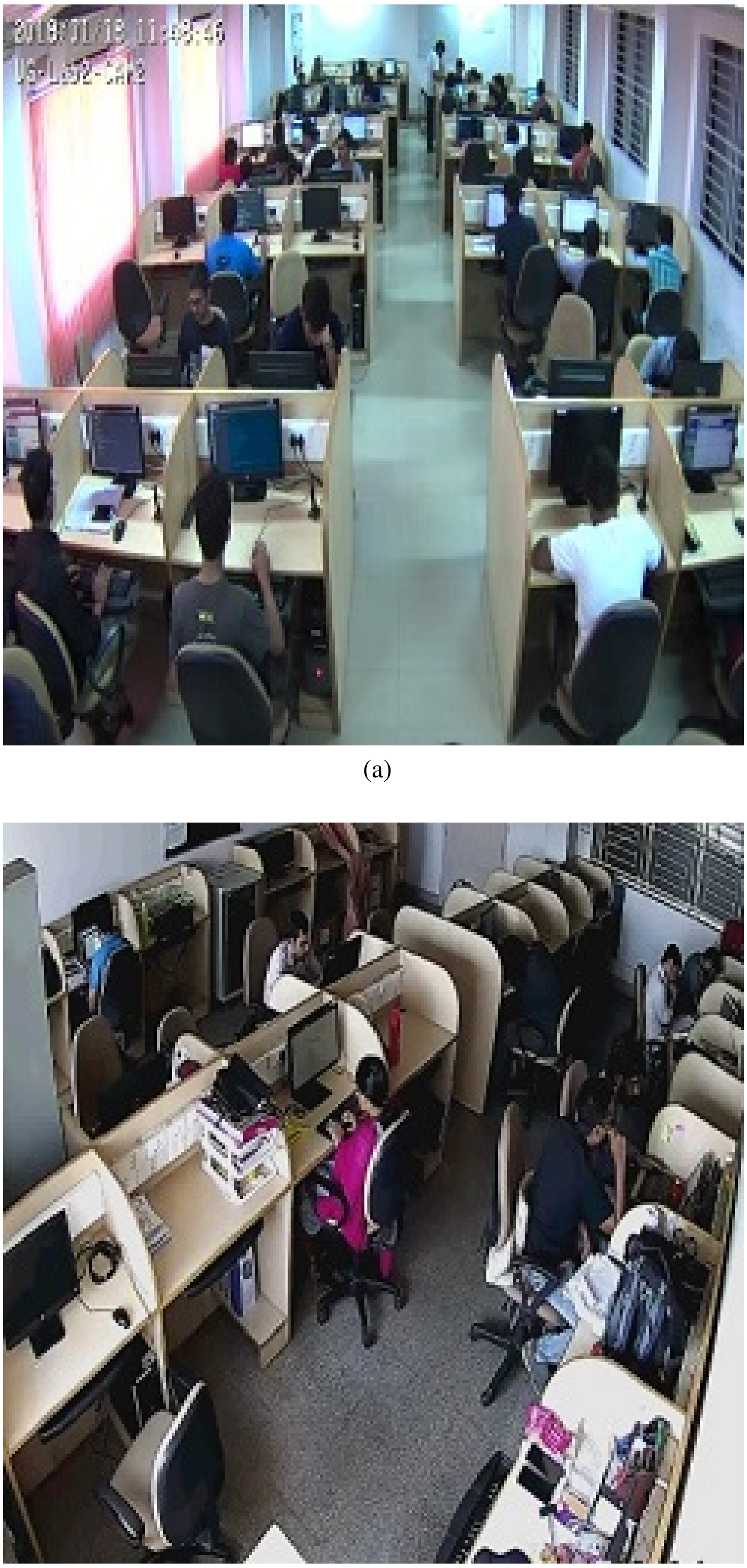
Action Label	**No. of Samples**
Discussion	5034
Engaged	40052
Sleeping	6916
Eating	2092
Using Smart Phone	768

total number of samples collected for the various action labels applied to the images is presented in Table 3.3.

The dataset was constructed from spontaneous videos. Multiple actions are determined from a single image frame. To circumvent potential dataset biases the annotation is done by three distinct annotators. The domain-specific expertise of the annotators was diverse. Also, Kappa Coefficient (Cohen, 1960) is utilized to determine the level of agreement between the annotators.

3.4 Experiments, Results, and Analysis

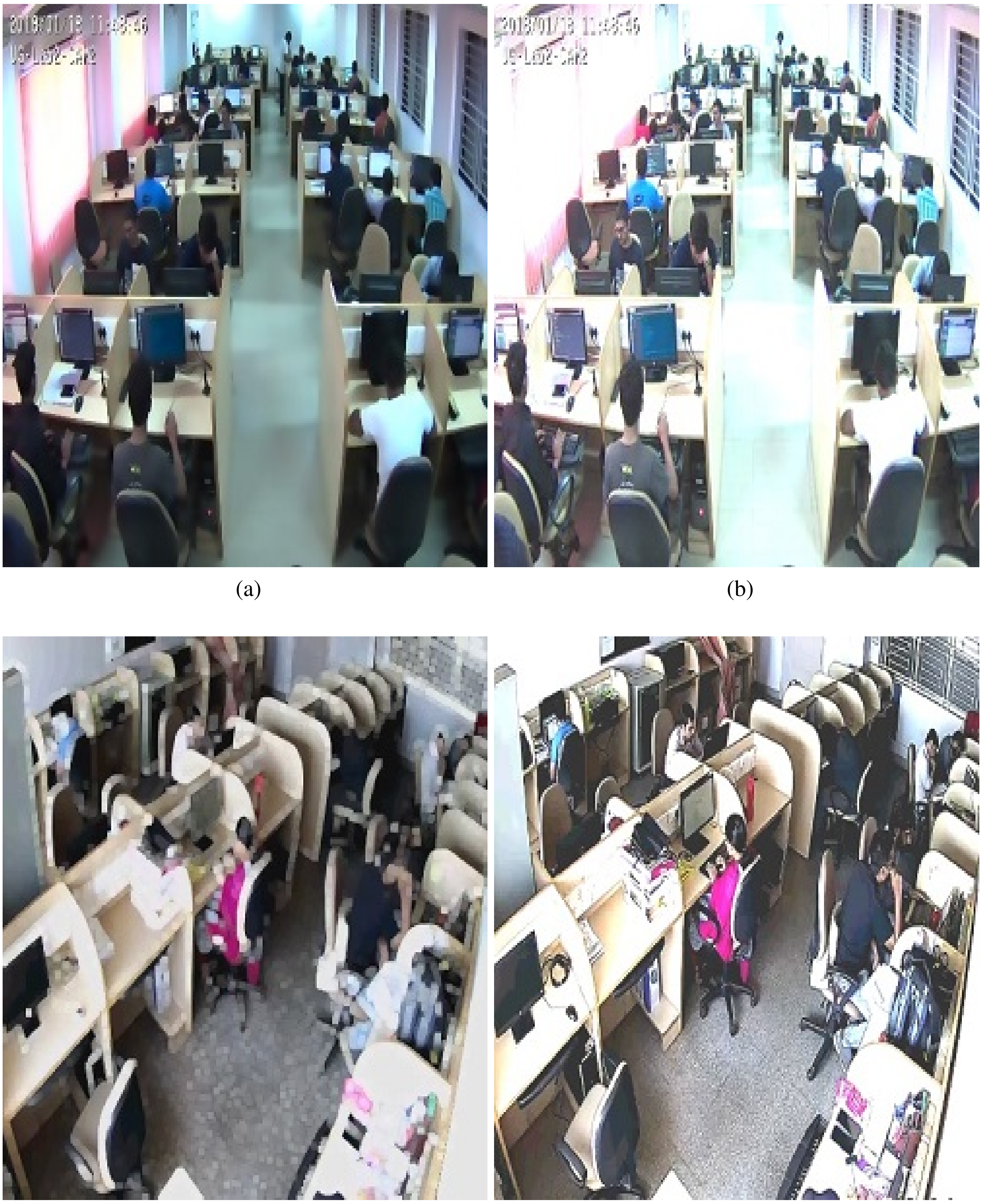
This section describes the specific observations made during the training and testing of the proposed model. The proposed model is trained utilizing the K-fold cross-validation method. The entire dataset is divided into ten folds, with 650 images in each fold. Each fold is tested, while the remaining nine are utilized for training.



(a)

(b)

Figure 3.3: Sample of original images used in STUDENT_ACTION dataset.



(a) (b)

(c) (d)

Figure 3.4: Sample of augmented images in STUDENT_ACTION dataset.

3.4.1 Testing

Figure 3.5 shows the ground-truth information of the set of frames in one of the fold used for testing. This testing fold in the best model has 650 frames with a total of 3984 *Engaged*, 511 *Discussion*, 706 *Sleeping*, 218 *Eating*, and 64 *Using_Smart_Phones*.

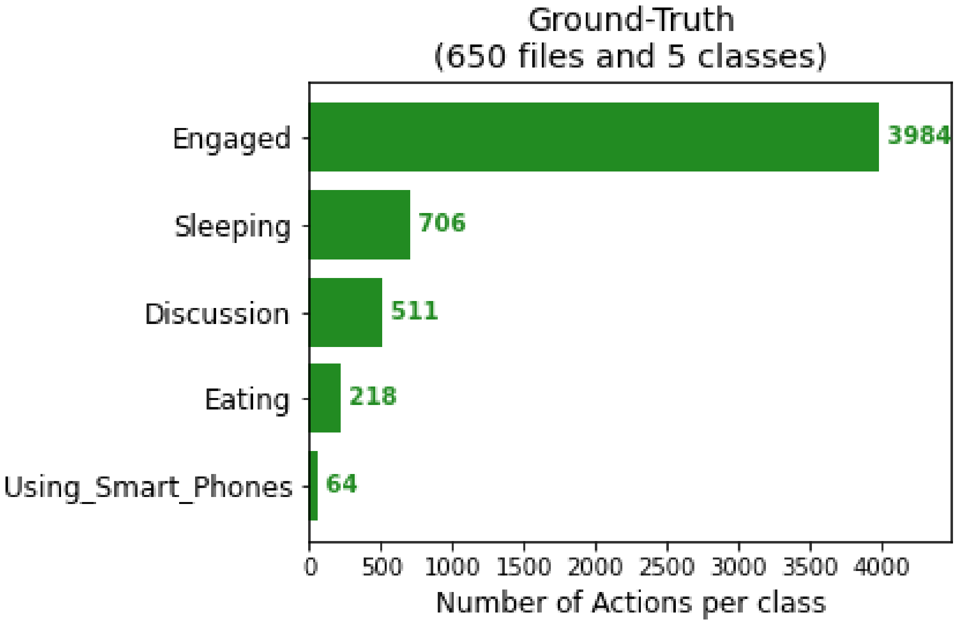


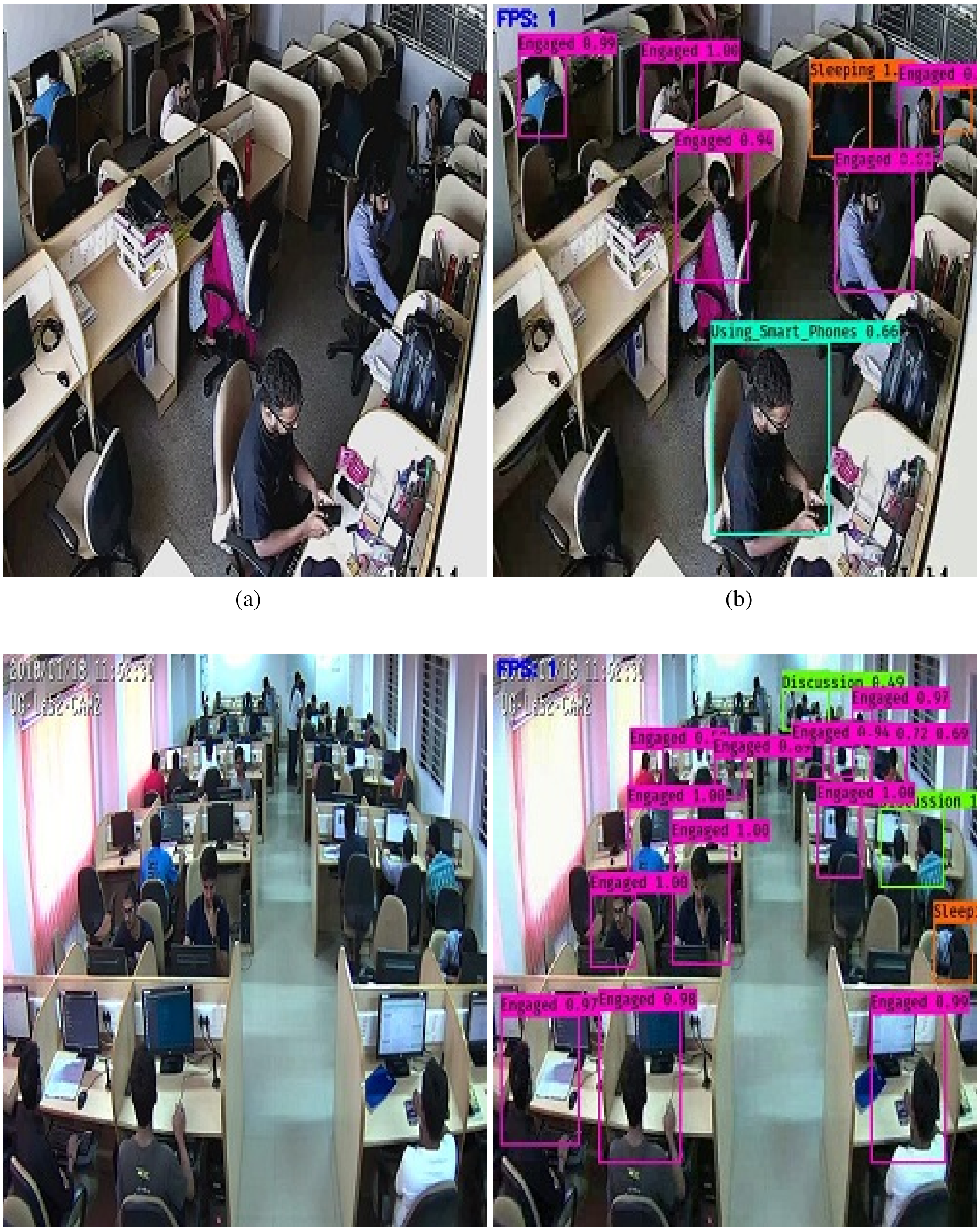
Figure 3.5: Action sample distribution in one of the fold in $10 - fold - Cross - Validation$

Figure 3.6 depicts the two sample input test images captured spontaneously and the corresponding output obtained from the proposed system. Here, the students are performing several actions inside the computer laboratory. The proposed system detected several *Engaged*, two *Sleeping*, and two *Discussion* actions. The output image displays the detected actions along with the score of action detection.

3.4.1.1 Intersection over Union

We also considered IoU, which stands for intersection over union, when evaluating the system. The proposed system must recognize and localize all actions within a single image frame. Intersection over Union (IoU) is an evaluation metric used to determine the correctness of an object detector. For the IoU measurement process, we consider both the ground truth bounding boxes and the predicted bounding boxes. The IoU can be mathematically described as Equation (3.2).

$$IoU = \frac{Area\ of\ Overlap\ of\ Two\ Bounding\ Boxes}{Area\ of\ Union\ of\ Two\ Bounding\ Boxes} \tag{3.2}$$



(a) (b)

(c) (d)

Figure 3.6: Sample input and the corresponding output images.

3.4.1.2 Kappa Coefficient

To gauge inter-rater agreement, (Cohen, 1960) first proposed using a Kappa coefficient. In the proposed system, the Kappa coefficient was used to evaluate the degree to which annotators agreed upon identifying and localizing the actions depicted in the dataset's image frames. The Kappa statistic can take on any value from 0 to 1. In this case, a value of 1 indicates that all annotators agree, while a value of 0 indicates no consensus. The Kappa coefficient is computed using Equation 3.3 based on (E, 2020):

$$K = \frac{N\sum_{i=1}^{n} m_{i,i} - \sum_{i=1}^{n} G_i * C_i}{N^2 - \sum_{i=1}^{n} G_i * C_i} \tag{3.3}$$

Where, G_i and C_i indicate true values and predicted values belonging to class i. Variable m indicates the confusion matrix.

i is the total number of action labels.

N is the total maximum of the total number of annotations by both annotators. $m_{i,i}$ is the number of values both annotators annotated as action label i.

C_i is the number of values belonging to action label i according to Annotator1.

G_i is the number of values belonging to action label i according to Annotator2.

Table 3.4: Interpretation of Kappa statistic

Kappa Value	**Level of Agreement**
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

In the proposed methodology for comparing annotations by different annotators, initially, C and G are considered to be Annotator1 and Annotator2 annotations, respectively. For Annotator3, C represents the mean of annotators 1 and 2, and G represents Annotator3. The Kappa Coefficient of annotations was found to be 0.65 according to the findings. The interpretation of various Kappa values is shown in Table 3.4 based on (Landis and Koch, 1977). Here, the range of 0.61 to 0.80 indicates that the two observers are in substantial agreement. The obtained Kappa Coefficient value is 0.65, indicating substantial agreement among the annotators and that the dataset is less biased, albeit not a perfect one.

3.4.1.3 Analysis of Detected Actions

Figure 3.7 depicts the detection result of the proposed system with IoU = 0.45. Table 3.5 details the True Positive (TP) and False Positive (FP) results for each action label obtained during testing.

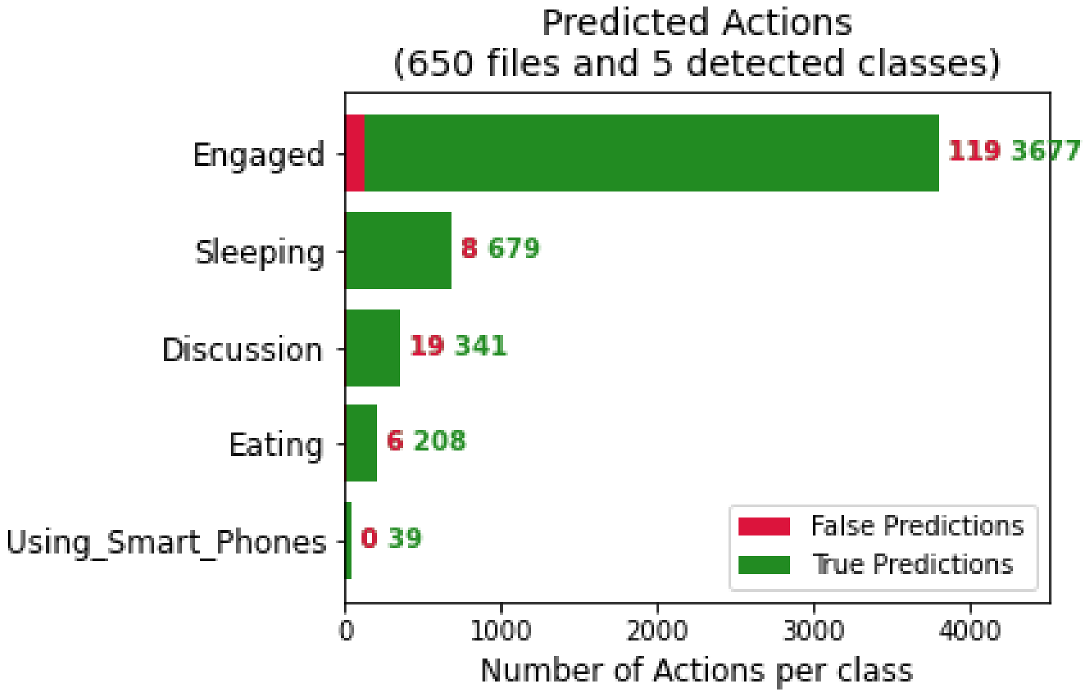


Figure 3.7: Information about detected actions.

Table 3.5: TP and FP found for action labels while testing.

Action Label	**Ground_Truth**	**Total Detected**	**True Positive**	**False Positive**
Discussion	511	360	341	19
Engaged	3984	3796	3677	119
Sleeping	706	687	679	8
Eating	218	214	208	6
Using_ Smart_Phones	64	39	39	0

Precision/Recall Curves:

Precision and recall are the two key metrics used to evaluate classifier's effectiveness. Precision is the proportion of relevant instances within the total number of instances retrieved. Recall is the proportion of retrieved relevant instances in relation to the total number of relevant instances. The Precision and Recall are computed as shown in Equations (3.4) and (3.5), respectively.

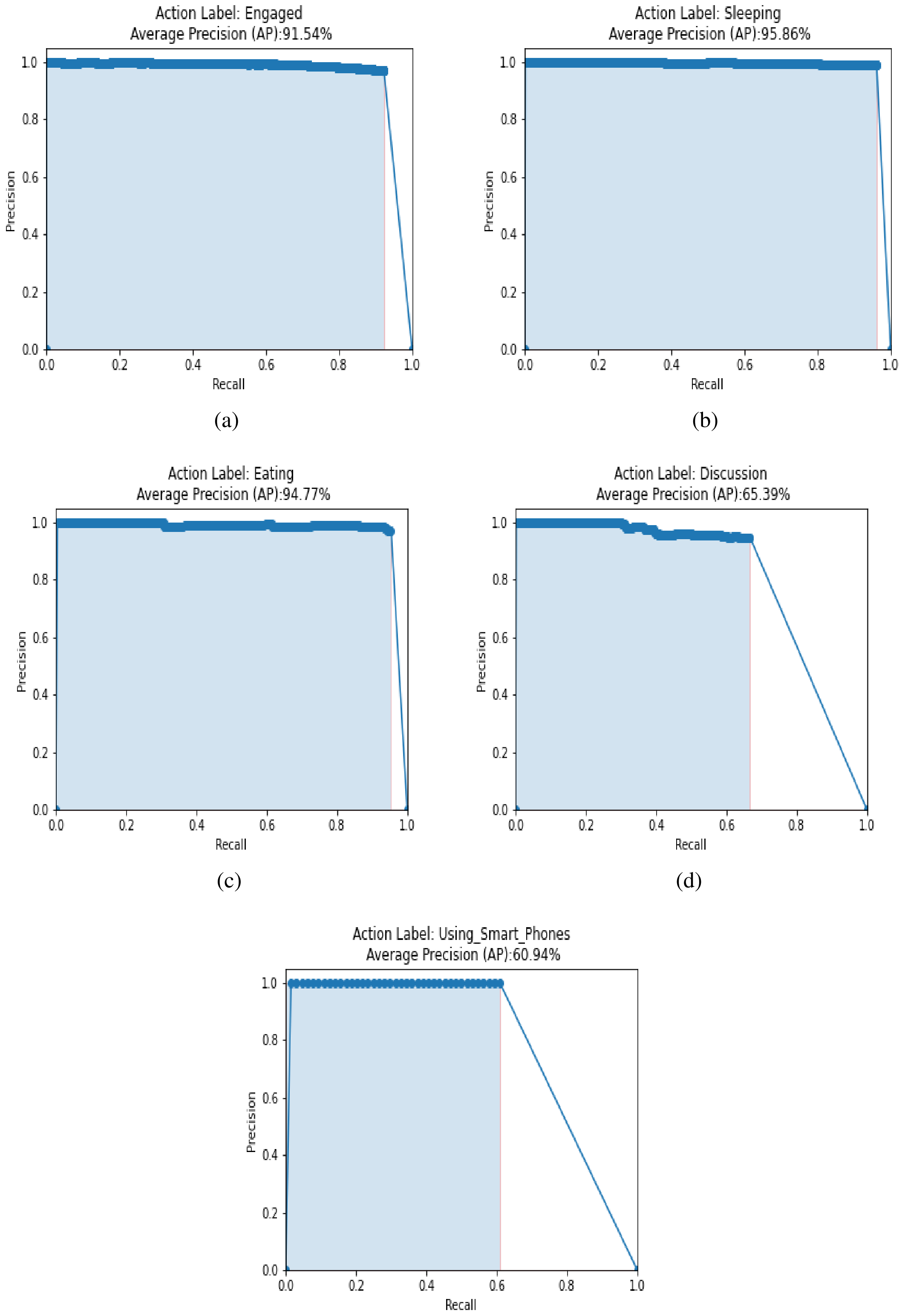
$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{3.4}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{3.5}$$

Precision-Recall (PR) curves provide a more accurate depiction of an algorithm's performance with highly skewed datasets (Davis and Goadrich, 2006). It summarises the trade-off between the true positive rate and the positive predictive value for a predictive model employing different probability thresholds, and it is preferable to use this in cases of moderate to substantial class imbalance (Brownlee, 2018). As in (Cartucho, 2018) and (Cartucho et al., 2018), we obtain PR curves for all five actions by mapping each detection to a ground-truth class instance. Figure 3.8 depicts the various PR curves obtained during testing for the five distinct actions. The average precision for *Engaged*, *Sleeping*, *Eating*, *Discussion*, and *Using_Smart_Phones* is 91.54%, 95.86%, 94.77%, 65.39%, and 60.94%, respectively. Using Average Precision (AP) for all action classes and the mean Average Precision (mAP), the performance of the proposed model is determined. Figure 3.9 depicts the AP and mAP obtained during testing for each action. The mAP determined through testing is 81.70%.

Finally, we used template matching in the action recognition subsystem to process the video in a timely manner. Algorithm 3.1 outlines the procedure that needs to be followed. Numerous experiments led to an empirical determination that a threshold of 0.97 was optimal for template matching. In cases where the current frame is very similar to the keyframe, the current frame is ignored, and only the actions from the keyframes are used.

Table 3.6 displays the evaluation results of three videos captured by CCTV cameras. The outcome demonstrates that template matching is faster than action detection. When comparing keyframes and skipped frames, there is also a small difference in the number



(a) (b) (c) (d) (e)

Figure 3.8: Precision/recall curves obtained for different actions.

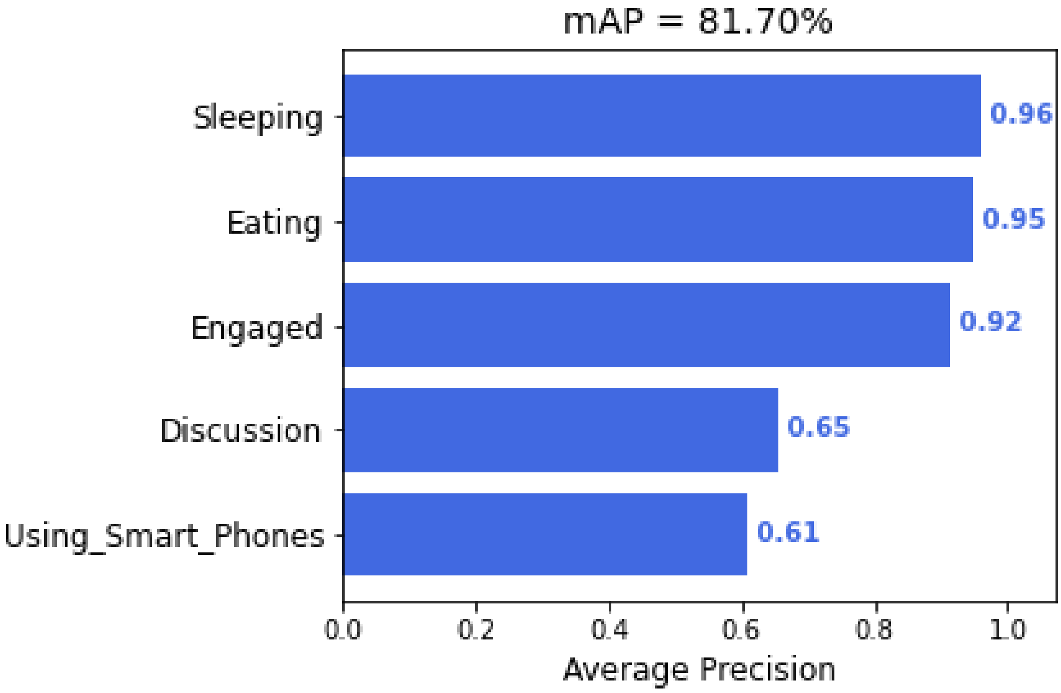


Figure 3.9: mAP obtained during testing.

of added or removed actions. The impact of cut or added actions is also minimal in the video. As a result, it drastically cuts down on time necessary to process the entire video. Consequently, by omitting extraneous frames, the final size of the stored video can be drastically decreased. Information about the time it took to process the full video with and without template matching on the computer system we used for testing, as well as the percentage of other actions, added and actions missed, can be found in Table 3.6.

The experiments showed that the time required for template matching in keyframe selection is very less (0.03 seconds per frame). But for action recognition, it was quite high (0.9 seconds per frame). The experiments showed that few actions are added or deleted while applying the keyframe selection. However, when the entire video is considered, the variations in the number of actions appear negligible. No more than 3% of total activities were changed between the first two keyframes. So, regarding both speed and accuracy, the proposed system demonstrated encouraging real-time performance.

Table 3.6: Comparison of video processing with and without template matching.

Name	**T1.mp4**	**T2.mp4**	**T3.mp4**
Total number of frames	500	650	1001
Number of skipped frames with template matching	429	298	982
Number of frames considered for action detection with template matching	71	352	19
Average time for action detection per frame (seconds)	0.9	0.9	0.9
Average time for template matching per frame (seconds)	0.03	0.03	0.08
Total time taken to process video with template matching (seconds)	84.30	352.59	108.18
Total time taken to process video without template matching (seconds)	464.35	610.86	955.41
Extra actions added w.r.t first keyframe (%)	0	1	0
Extra actions added w.r.t second keyframe (%)	3	0	0
Actions deleted w.r.t first keyframe (%)	3	0	2.04
Actions deleted w.r.t second keyframe (%)	1.5	0	0

3.4.2 Limitations of the Work

Although the proposed system achieves promising results in detecting and localizing student's actions, the proposed model can recognize a few actions correctly. Within the various categories, there were instances of improper classification. In addition, it has been discovered that the model can efficiently identify and localize specific types of actions. The imbalance impacts the proposed model's performance in the dataset. Further, the proposed model can recognize only simple actions and cannot recognize complex actions that involve temporal cues.

3.5 Summary

A potential solution for a real-time student action recognition system based on transfer learning and YOLOv3 is proposed in this work. The proposed model enables the recognition and localization of students' actions within the computer laboratory on the university campus, and the results are encouraging. The model also demonstrates that a specific set of human actions can be recognized from a single image frame. By reducing the total number of frames that needed to be processed, the proposed model was able to reduce the amount of time required to perform video analysis. Consequently, the proposed system can be utilized for real-time monitoring of computer laboratory activities. The RGB data is computationally intensive.

The initial model was trained to identify objects. Therefore, there is a domain shift when it is used in the laboratory for target action recognition tasks. So, transfer learning is applied by loading the weights of the original deep learning model trained to detect objects and modifying the classification layer. In addition, action recognition is carried out in still images using only spatial features. Further, the proposed models' robustness in recognizing actions in other domains can be verified by constructing similar domain-specific action datasets.

The availability of depth sensors provides data in different modalities like depth and skeleton. Based on the literature review, the skeleton and depth require fewer resources. The next chapter discusses the skeleton data based HAR for the single-view scenario.

Chapter 4

Single-view Human Action Recognition System using Skeleton Data

An action is defined by a series of human body movements that involve multiple body parts simultaneously. Human Activity/Action Recognition (HAR) plays a crucial role in a vast array of applications, such as smart video surveillance (Liu et al., 2018), home monitoring systems (Foroughi et al., 2008), intelligent human-machine interfaces (Ramezani and Yaghmaee, 2016), and many others. Due to the complexity of human body movement, while performing an action, HAR is one of the most challenging research topics in computer vision. The way in which a person's body moves while performing the same action varies from person to person. In addition, it is challenging to accurately represent the spatio-temporal features of actions when various environmental factors, such as the lighting condition, are considered (Nie et al., 2019).

As inexpensive depth sensors like Microsoft Kinect have become more widely available, the rate at which new skeleton data is being produced has accelerated (Firman, 2016). Instead of storing each video frame, the joint coordinates of the human body are stored in skeleton data. Therefore, much less space is required to store skeleton data when compared to other data modalities. The coordinates remain unchanged regardless of the observer's position, making skeleton data stable across viewpoint changes (Nguyen et al., 2018). Due to its robustness to changes in the background, such as lighting, clothing condition, and more efficient computation, skeleton-based action identification gained popularity (Fan et al., 2020).

The human skeleton is composed of numerous joints. This thesis work represents the skeleton data as a graph consisting of skeleton joints as vertices and connections between these vertices as edges. Since the human skeleton data is represented as a graph, the various graph geometries can be considered for the target task using skeleton (Wang and Wang, 2018). Geometries can include distance, surface area, angle, etc. The distance between human skeleton joints varies depending on the action being performed. Likewise, the angle between the joints varies depending on the action. This thesis work focused on the distance and angle between 3D coordinates of skeleton joints for HAR.

The proposed model attempted to reduce the computation cost without degrading the performance by limiting feature extraction to the joints involved in actions.

Most informative joints provide crucial data for action recognition in the skeleton model. For this reason, this work considered the most informative distance and the angle between the joints when building the feature set, as suggested by (Nguyen et al., 2018). Recently, DL models have shown promising results in classification tasks. Thus, we proposed a DL model consisting of Dense and Softmax layers for HAR. The significant contributions of this thesis work are listed below.

Contributions:

- A new tree representation of skeleton joints in skeleton data.
- Efficient feature extraction method to extract most informative distance and angle features from new representation of skeleton data.
- A deep learning model with Dense and Softmax layers followed by score fusion method to improve overall performance of HAR system while considering the complete video.

4.1 Proposed Methodology

Figure 4.1 depicts the proposed architecture for the action recognition task. The proposed system comprises three subsystems: Feature extraction and fusion, the DL model, and Score fusion. Given a skeleton sequence of an action video with N frames, where each frame has m 3D skeleton joint coordinates $fi = j1, j2, ...jm$, the proposed framework extracts features from skeleton joint information from each skeleton frame of the action video. The newly developed neural network model is trained with extracted features and generates a score of action recognition for each frame in the action video. The score fusion model recognizes the final action in the video by combining the scores in each skeleton frame in an action video. Detailed descriptions of each step are as follows.

4.1.1 Feature Extraction and Feature Fusion

Most skeleton-based human action recognition systems extract features using information from all joints (Wint Cho et al., 2018). However, the degree to which each joint is

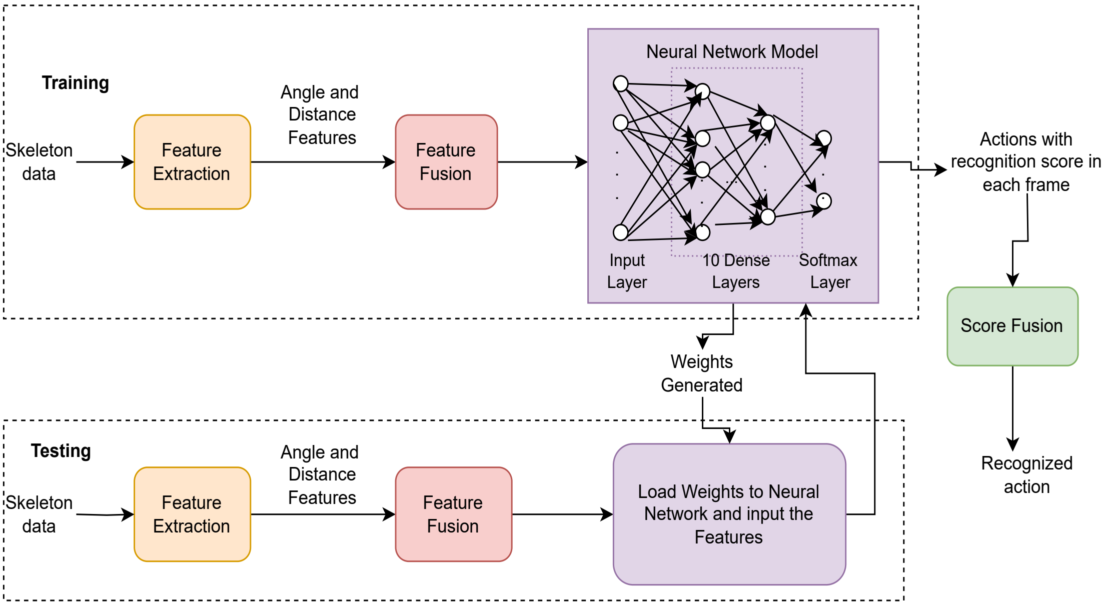


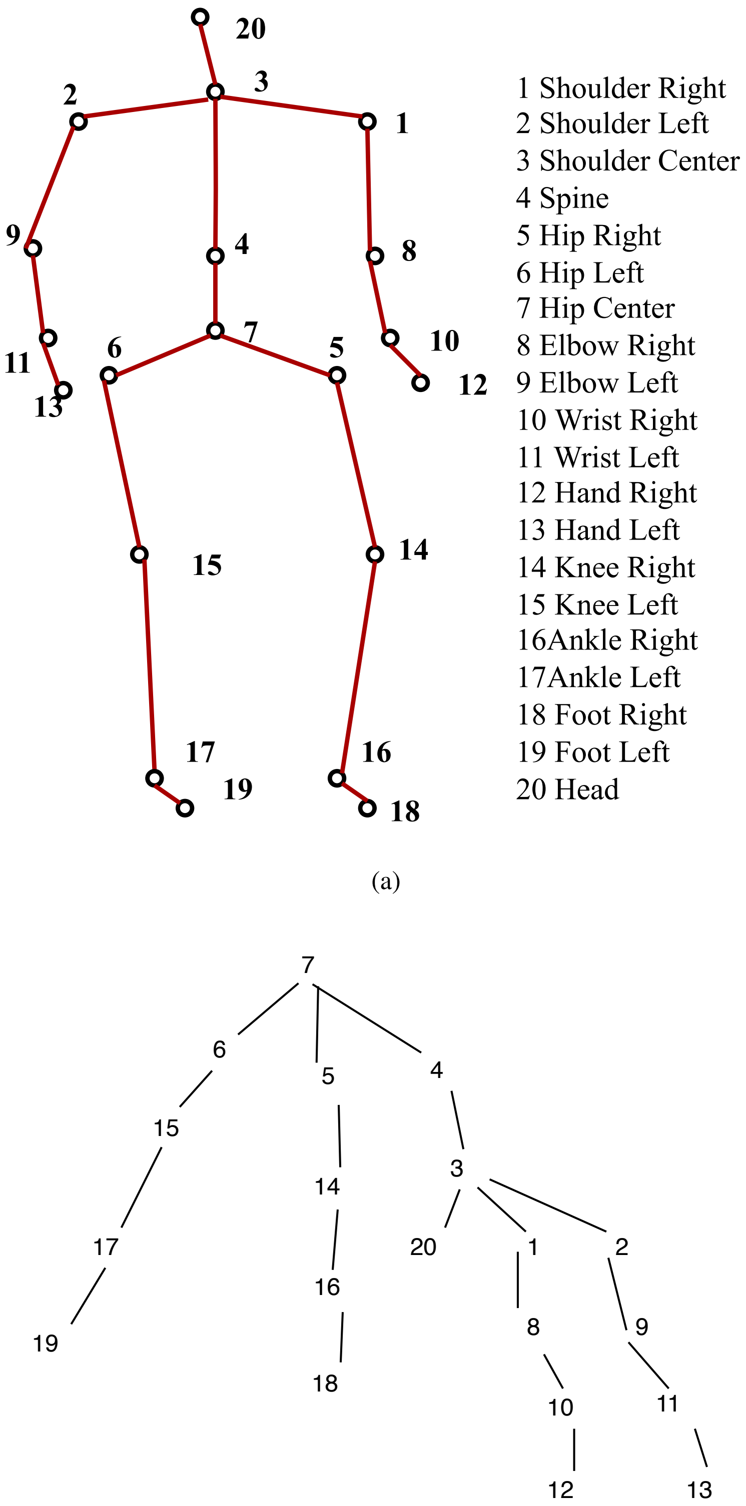
Figure 4.1: Proposed framework for 3D human action recognition using skeleton data.

involved in an action depends on the action being performed. Depending on the action performed, various body joints will contribute in various ways. For feature extraction, the proposed method utilized the most frequently used skeleton joints during action execution. To accomplish this, we proposed a new representation of the skeleton joints in a skeleton frame. The skeleton's joints are depicted as a tree, with the hip-center joint serving as the tree's root. The subtrees of the tree are then determined by traversing the joints from the hip-center joint using the Depth First Search (DFS) method. The skeleton joint data obtained from Kinect v1 depth sensor and its proposed tree representation of joints are depicted in Figures 4.2 (a) and 4.2 (b), respectively.

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}, \tag{4.1}$$

$$\Theta = cos^{-1}\left(\frac{\vec{AB}.\vec{CB}}{\|\vec{AB}\|\|\vec{CB}\|}\right), \tag{4.2}$$

A significant correlation exists between the action and the distance and angle between the joints of a skeleton model. This thesis work computes the distance between every other pair of 3D joints, ignoring the joint that comes immediately after the joint under consideration. In addition, the proposed method uses the angles that provide the



(a)

(b)

Figure 4.2: Original skeleton format and new representation. (a) Skeleton joint data from Microsoft Kinect v1. (b) New tree representation of skeleton joints.

most information between the 3D joints. The Euclidean distance between any two 3D joint coordinates $A = (x_1, y_1, z_1)$ and $B = (x_2, y_2, z_2)$ is calculated using the Equation (4.1). Similarly, given three 3D joint coordinates $A = (x_1, y_1, z_1)$, $B = (x_2, y_2, z_2)$, and $C = (x_3, y_3, z_3)$, the angle $\Theta = \angle ABC$ is computed using Equation (4.2). After the distance and angle features have been extracted, they are combined to create a feature vector with a size of 36 for each frame in the action video.

4.1.2 Neural Network Model

The proposed deep learning model with Dense layer and ReLU activation function, and Softmax layers is trained for action recognition using the extracted features. The proposed model used Adam optimizer with learning rate of 0.001 during training. Based on its computation, the model assigns a probability score to each action in each frame.

4.1.3 Score Fusion

This subsystem considers each frame's probability score of recognition in an action video and outputs an action that has been recognized. For each frame, the proposed model considers the top three scores and the action labels associated with them. Later the sum of the score for each action label is calculated. The action with the highest sum is considered a recognized action from the video.

Algorithm 4.1: 3D HAR System using Skeleton Data.

```
Input: Skeleton sequence of each frame in a action video
Output: Action label of recognized human action
while not end of frames in action video do
    Extract 3D joint coordinates of the skeleton found in an image frame.
    Arrange the joints coordinates in a tree representation.
    Extract the required distance and angle features.
    Input features into a neural network to generate a score of each action
      recognition.
end
for each skeleton with recognition score do
    Sort the recognition score in the descending order.
    Select the first three scores and corresponding actions in each frame say
      $s_1, s_2, s_3$ and $a_i, a_j, a_k$.
end
Find the sum of scores for each action $a_l$ where $l = 1$ to $N$ where, $N$ is the
  number of actions by considering all image frames in action video.
Select the action with the highest sum of the score as the recognized action.
```

Algorithm 4.1 depicts the overall procedure of the proposed HAR system using skeleton data. Algorithm 4.1 shows the main steps involved in the overall procedure including feature extraction to action classification.

4.2 Experiments, Results, and Analysis

4.2.1 Dataset

The proposed system is evaluated using the MSRAction3D benchmark dataset for 3D human action recognition, which is publicly available (Li et al., 2010). Due to the similarity of many actions, it is among the more challenging datasets for action recognition (Li et al., 2016). The dataset contains human actions captured using a Kinect depth sensor. There are 20 human actions in the dataset, performed twice or thrice by ten subjects. It provides skeleton data in the form of twenty joint coordinates. There are 567 action sequences in total, but the ten with the most background noise were filtered out. The proposed system will incorporate the remaining 557 sequences.

Table 4.1: Three groups of actions AS1, AS2, and AS3.

AS1	**AS2**	**AS3**
horizontal arm wave	high arm wave	high throw
hammer	hand catch	forward kick
forward punch	draw x	side kick
high throw	draw tick	jogging
hand clap	draw circle	tennis swing
bend	two hand wave	tennis serve
tennis serve	forward kick	golf swing
pickup and throw	side boxing	pickup and throw

The actions in the dataset are organized into three overlapping sub-sets (groups) of eight classes belonging to the action groups, namely: AS1, AS2, and AS3. Each group's action labels are shown in Table 4.1. AS1 and AS2 categorize actions that share similar motions, while AS3 contains actions with more complex movements. Several evaluation protocols are suggested in the literature for evaluation of HAR system on MSRAction3D dataset. Two tests are conducted to evaluate the proposed system's performance, each utilizing the evaluation protocols described in (Li et al., 2010; Presti and La Cascia, 2016). The evaluation protocols are listed below.

- Protocol P-1: 3-Fold Cross-validation with 1/3 of the data is used for testing and the remaining 2/3 for training.
- Protocol P-2: 3-Fold Cross-validation with 2/3 of the data is used for testing and the remaining 1/3 for training.

4.2.2 Experiments and Results

The proposed systems' performance is evaluated on three action groups AS1, AS2, AS3, and the entire dataset, using both evaluation protocols mentioned earlier. The sample distribution among testing and training is as shown in Table 4.2.

Table 4.2: Data sample distribution among testing and training in P-1 and P-2.

		AS1	**AS2**	**AS3**	**Entire Dataset**
P-1	**Training**	146	152	147	371
	Testing	73	77	74	186
P-2	**Training**	73	77	74	186
	Testing	146	152	147	371

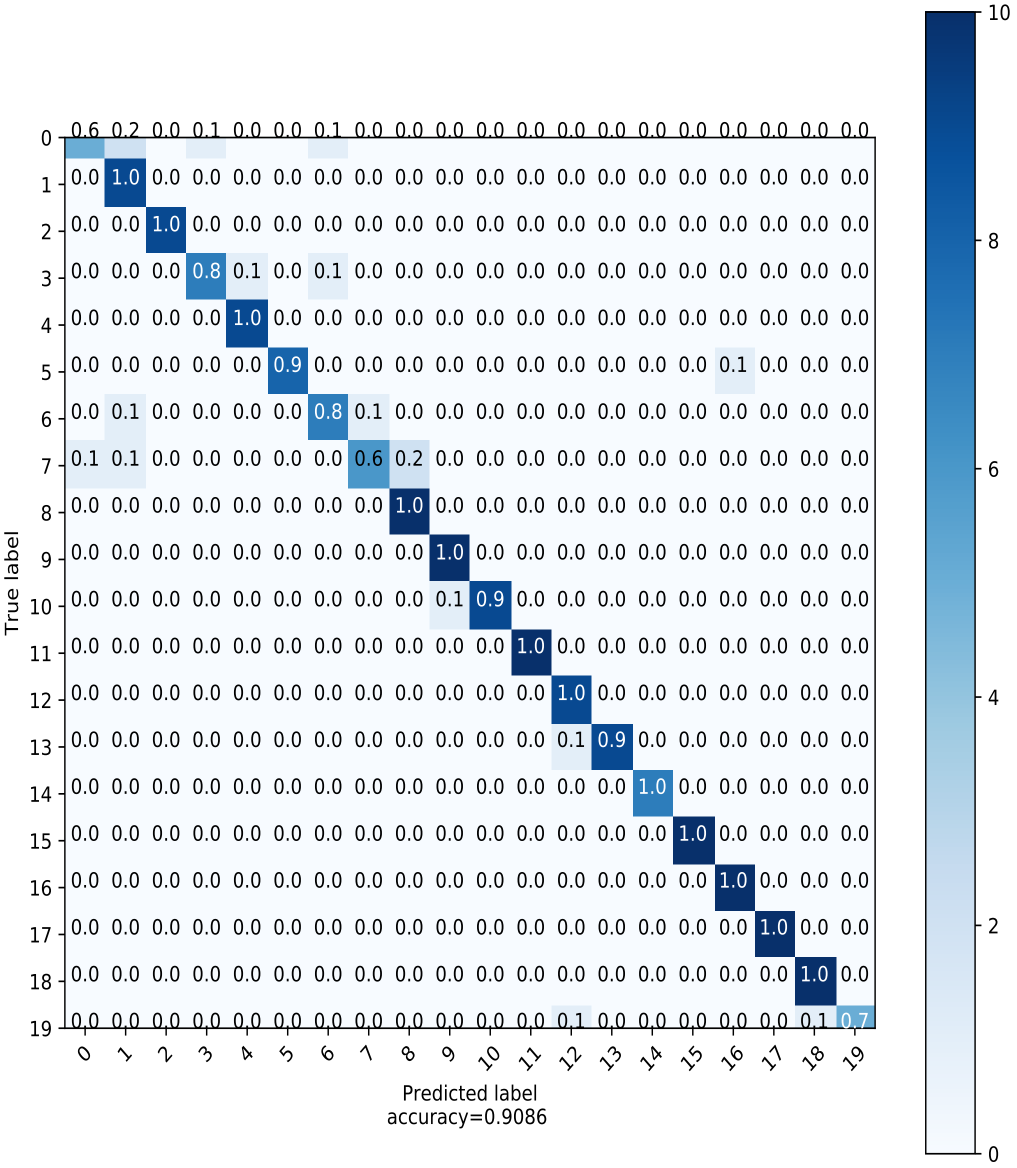
The proposed neural network model was trained using two distinct methods for both evaluations. The first strategy (M-1) divides the training data into a training dataset and a validation dataset with a ratio of 8:2 between the two. Further, the developed model was evaluated using the dataset's test subset. For the second method (M-2), we employed the K-fold cross-validation (James et al., 2013) technique. Here, all ten folds are validated by splitting the training data into ten subsets (folds) and training the model on nine folds. On the test segment of the dataset, the constructed model is evaluated.

The performance of proposed method on MSRAction3D dataset using both evaluation protocols is reported in Table 4.3. The results show that with the entire dataset taken into account, the proposed model achieves an accuracy of 90.86% using the method M-2 and evaluation P-1. On AS1, using P-1 we achieved the best Accuracy of 95.83%, and P-2 achieved 95.17% Accuracy. On AS2, P-1 and P-2 achieved the best Accuracy of 89.61% and 87.58%, respectively. On AS3, P-1 and P-2 achieved the best Accuracy of 98.63% and 96.62%, respectively.

Table 4.3: Action recognition Accuracy of the proposed model for HAR using skeleton data.

Data	**Method**	**Protocol**	**Accuracy (%)**
Entire Dataset	M-1	P-1	87.09
		P-2	85.12
	M-2	P-1	90.86
		P-2	88.40
AS1	M-1	P-1	95.83
		P-2	91.72
	M-2	P-1	95.83
		P-2	95.17
AS2	M-1	P-1	84.92
		P-2	84.96
	M-2	P-1	89.61
		P-2	87.58
AS3	M-1	P-1	94.52
		P-2	95.94
	M-2	P-1	98.63
		P-2	96.62

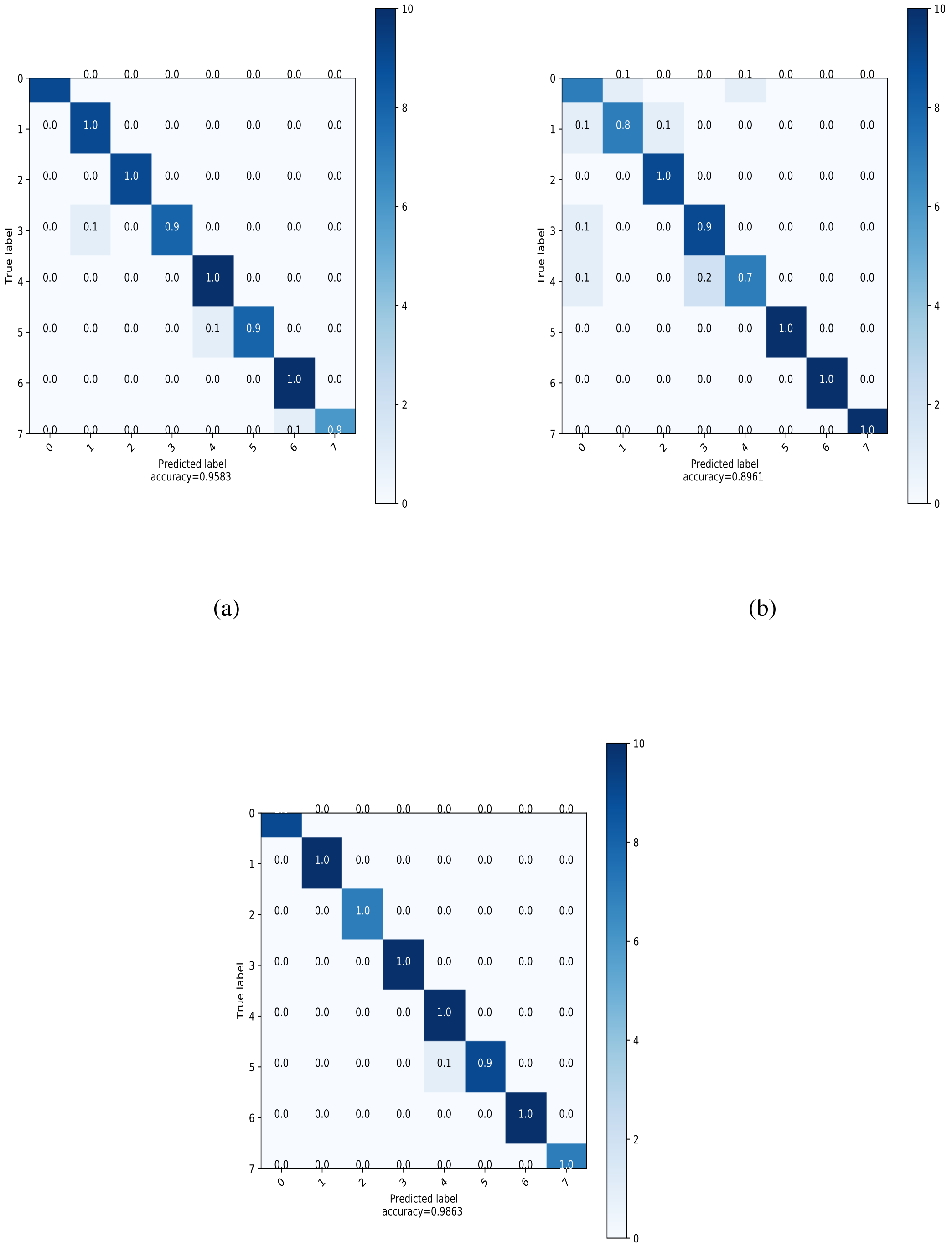
Figure 4.3 depicts the confusion matrix obtained while considering the entire dataset with P-1. The labels 0 to 19 denote the identifiers for action labels. It shows that 15 actions were recognized with more than 90% Accuracy. The overall accuracy obtained is 90.86%. Figure 4.4 depicts the confusion matrices for the best results of AS1, AS2, and AS3, respectively. For AS1, the Figure 4.4 shows the results of P-1, in which all the actions are recognized with above 90% Accuracy. It also demonstrates that, in AS2, five actions were recognized with above 90% Accuracy, and two actions with above 80% Accuracy using evaluation protocol P-1. In AS3, seven actions were recognized with 100% Accuracy and one with 90% Accuracy using evaluation protocol P-1.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0	0.6	0.2	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.8	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0
6	0.0	0.1	0.0	0.0	0.0	0.0	0.8	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.6	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.9	0.0	0.0	0.0	0.0	0.0	0.0
14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
15	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
17	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
18	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
19	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.7

Figure 4.3: Confusion matrix using evaluation protocol P-1 for the entire dataset.

Further, the performance of proposed system is also evaluated using $Precision$, $Recall$, and $F - Score$ measures. Table 4.5 reports the average $Precision$, average $Recall$, and $F - Score$ obtained from the proposed model over AS1, AS2, AS3,



(a)

(b)

(c)

Figure 4.4: Confusion matrices. (a) AS1. (b) AS2. (c) AS3.

and entire dataset with best combination of training approach and evaluation protocol. $Precision$ and $Recall$ for each label in each set of data are calculated using Equations (3.4), and (3.5), respectively. $F - Score$ is computed using Equation (4.3). The proposed model demonstrated the highest $Avg(Precision)$, $Avg(Recall)$, $F - Score$ of 98.86%, 98.75%, 98.80%, respectively over set AS3 on protocol P-1 with M-1.

$$F - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4.3}$$

The $Avg(Precision)$ and $Avg(Recall)$ are computed using Equations (4.4) and (4.5), respectively. Where, N, and i denotes the number of action labels, and the action label, respectively.

$$Avg(Precision) = \frac{\sum_{i=1}^{N} Precision_i}{N} \tag{4.4}$$

$$Avg(Recall) = \frac{\sum_{i=1}^{N} Recall_i}{N} \tag{4.5}$$

Table 4.4: $Precision$, $Recall$ and $F - Score$ obtained for HAR using skeleton data.

Evaluation Metrics	**AS1**	**AS2**	**AS3**	**Entire Dataset**
Average $Precision$ (in %)	97.61	89.60	98.86	85.14
Average $Recall$ (in %)	95.43	89.44	98.75	87.84
$F - Score$ (in %)	96.5	89.51	98.80	86.47

While considering the most informative joints, the proposed approach used the general perspective. Though there are variations in the speed at which the body part moves during the action, there are similarities in how body parts move. The proposed approach performs frame-wise classification of the actions and then performs score fusion to find the action to which the maximum number of frames belong in an action video to compensate for the difference in people acting and make it generalized across individuals.

The experimentation shows that, though the proposed approach showed robust performance in recognizing a greater number of actions, there are few exceptions. For example, considering Figure 4.3 the confusion matrix obtained considering the entire dataset, the proposed approach is robust in recognizing 12 different types of actions accurately irrespective of the individuals performing the action.

4.2.3 Comparison with Existing Works

The proposed systems' performance is compared with the baseline methods using the same evaluation protocols on AS1, AS2, and AS3 action subsets of MSRAction3D. Table 4.5 reports the comparison results. The proposed work's performance is better than those obtained by (Li et al., 2010). The proposed model achieved an average accuracy of 94.69% when using P-1 and 93.12% when using P-2.

Table 4.5: Comparison of Accuracy (in %) of the proposed work with baseline methods.

	Protocol	**AS1**	**AS2**	**AS3**	**Average**
(Li et al., 2010)	P-1	93.4	92.9	96.3	94.2
	P-2	89.5	89.0	96.3	91.6
(Yang and Tian, 2012)	P-1	97.3	98.7	97.3	97.7
	P-2	94.7	95.4	97.3	95.8
(Devanne et al., 2013)	P-1	93.4	93.9	98.9	95.4
	P-2	90.3	91	98	93.1
Proposed Method	P-1	95.83	89.61	98.63	94.69
	P-2	95.17	87.58	96.62	93.12

4.2.4 Computational Cost Analysis

The proposed method's computational cost analysis involves the complexity of feature extraction, DL model, and score fusion operation. For feature extraction, the majority of existing methods take into account all skeleton joints. For instance, if a method takes into account the distance between all 3D joints in a skeleton model with M joints, then the computational complexity for the feature extraction step is $O(M^2.N)$, where N is the number of frames in an action video. Using only the most informative joints, the proposed method constructs a 36-element feature vector from each frame of an action video. The estimated complexity of the feature extraction step is $O(N)$ times the number of features, where N is the number of frames in an action video. Thus, the cost of computation is drastically reduced. The complexity of DL model is approximately equal to the number of parameters in it. So, we can estimate the complexity of DL model as $O(W)$, where W denotes the number of parameters. Finally, the score fusion operation involves sorting the probability scores and picking top-3 actions from each frame. The estimated computations of this step will be $O(m \;\; log \;\; m)$, where m is the number of action labels in the dataset. Thus, the estimated complexity of overall procedure is as shown in Equation (4.6).

$$O(N) + O(W) + O(m \; log \; m) \tag{4.6}$$

4.3 Limitations

Even though the proposed method extracts frame-wise features and performs initial frame-wise classification, the total number of features for HAR is significantly reduced. However, the approach needs to be more efficient to capture the temporal aspects of human action effectively. Therefore, it is possible that the proposed method will not reveal the differences between the actions in an efficient manner if the actions involve more similar movements.

4.4 Summary

This chapter proposed a novel skeleton representation for efficient feature extraction. Further, a deep neural network model was developed for 3D HAR using the skeleton data. The proposed approach reduced the number of computations by constructing the feature set using only the most relevant joint information. The extracted features are then learned with the help of an DL model that contains Dense layers. Two different benchmark evaluation protocols are utilized to assess the proposed method's effectiveness on the benchmark publicly available 3D human action dataset MSRAction3D.

During this work it is found that the position of the body joints varies in the time domain as action is performed. Tracking this variation of features is essential in recognizing complex human actions. Based on the literature review, it is clear that the variants of RNN models are widely used to capture temporal aspects in addition to spatial features (Du et al., 2015; Jiang et al., 2020). Several researchers are working on extracting different kinds of temporal aspects for accurate action recognition. In longer-duration action, there is a dependency on how the body parts move in the time domain to accomplish the task. So, there is a need to capture the temporal aspects of human actions. Recognizing longer-duration actions accurately with the help of temporal aspects demand for more resources but will be useful in real-time scenarios. The next chapter gives a detailed explanation of HI system using 3D skeleton data-based gait-event specific features and DL models.

Chapter 5

Human Identification System using Single-/ Multi-view Skeleton-based Gait Data

In many embedded applications of intelligent surveillance systems in smart environments, reliable identification of humans at a distance without physical contact is gaining importance. These systems play a major role in many practical applications like forensics, terrorist monitoring, crime prevention, and access control (Huynh-The et al., 2020). Identifying humans with video surveillance systems is challenging for several reasons, including poor camera perspectives and environmental constraints. In surveillance systems, HI based on face recognition fails when the face is obscured by a mask, hand, or hat (Batchuluun et al., 2018; Sepas-Moghaddam and Etemad, 2023). Gait-based HI is gaining importance as a means of overcoming these challenges.

Gait is the series of coordinated, rhythmic movements that allow humans to move from one location to another (Boyd and Little, 2005). During gait recognition, we intend to assess human activity as a whole instead of analyzing individual body parts. Individuals' gait data can be collected at a distance without their knowledge or cooperation, making gait-based HI unobtrusive (Ye et al., 2020). (Kastaniotis et al., 2016) explained that the gait cycle comprises two phases: Stance and Swing, with any leg as a reference point. The Stance phase comprises sixty percent of the gait cycle, while the Swing phase accounts for the remaining forty percent. Multiple events occur during these gait phases. The gait cycle officially begins when the Stance phase's IC occurs, and the Swing phase's Terminal Swing concludes. Table 5.1 displays the various events and their relative proportion during a gait cycle based on (Webster and Darter, 2019).

This thesis work focuses on extracting gait event-specific features and developing novel DL models for HI based on these features. This work introduces two approaches for HI based on feature extraction and DL model. The developed systems are evaluated using benchmark evaluation protocols on both single and multi-view benchmark datasets using $Accuracy$, $Precision$, $Recall$, $F-Score$, and Cumulative Match Characteristic (CMC) curves.

Table 5.1: Gait cycle events and duration.

Gait Event	**Duration**
Loading Response	(0-10%)
Mid Stance	(10-30%)
Terminal Stance	(30-50%)
Pre Swing	(50-60%)
Initial Swing	(60-73%)
Mid Swing	(73-87%)
Terminal Swing	(87-100%)

For the most part, over the years, video-based methods have been proposed to aid in gait recognition. These video-based methods can be divided into two categories: model-free and model-based (Kumar et al., 2021). Model-free methods focus on directly extracting static and dynamic features from images of walking sequence. Most of these techniques obtain gait data by analyzing silhouette image-based human body features. Model-based approaches analyze the shape and kinematics of human body parts in order to recognize gait. Most of these model-based gait recognition techniques focused on human body joint distances or angles for feature extraction (Wan et al., 2018). Recent advancements in 3D visual depth sensors, such as Microsoft Kinect (Zhang, 2012), have led to an exponential increase in the number of 3D model-based approaches. The 3D skeleton joint data provided by the Kinect depth sensor obviates the need for complex procedures for building a model from visual data streams (Deng and Wang, 2019). In recent years, many skeleton-based gait recognition systems have been proposed (Choi et al., 2019; Khamsemanan et al., 2018; Deng and Wang, 2019; Bari and Gavrilova, 2019; Limcharoen et al., 2020). In addition, the success of DL (Goodfellow et al., 2016) models on image/video-based classification tasks prompted researchers to conduct new DL-based research on video-based HI systems.

Despite the vast literature on HI based on gait, the event-level features of gait for HI from walking have received insufficient attention. Because of issues like occlusion, differences in clothing, and the temporal aspects of various features while in motion, etc., gait recognition remains one of the most challenging tasks in computer vision. Here, we present two key works that utilize gait event-specific features and DL models for HI.

5.1 LSTM Model-based HI System using 3D Skeleton-based Gait Data

A method for minimizing the influence of noisy and occluded joints on gait recognition is proposed. In addition, a DL model trained on gait events for gait cycle classification, followed by a fusion operation for individual identification, is proposed in this study. The key contributions of this thesis work are listed below.

Contributions:

- Represented each gait event as a timestamp in the entire gait cycle. Also, proposed a set of gait event-level features to capture the spatial-temporal information of the gait cycle.

- Formed a quantitative summary of various features extracted from frames in each timestamp to minimize the effects of noise and the occlusions. As a result, the proposed HI system's total computational cost is significantly reduced.

- Proposed a DL model based on LSTM (Hochreiter and Schmidhuber, 1997) to efficiently recognize the gait cycle using features extracted from the sequence of gait events.

5.1.1 Proposed Methodology

The primary goal of the proposed system is to recognize a human from their walking pattern. Figure 5.1 shows the overall workflow of the proposed system for HI using skeleton based gait data. The overall workflow is separated into two distinct phases, known as the training and identification phases. During the training phase, a DL model trained to identify the gait cycles. The DL model developed during the training phase is used to obtain gait cycle-matching probability scores during the identification phase. The few initial subsystems within the training and identification phases share a common set of operations. Each phase begins with the detection and extraction of gait cycles from a person's walking sequence, followed by the extraction of features from individual gait events to produce a final feature vector representing the entire gait cycle. The extracted features are then used to train the proposed LSTM-based DL model to identify gait cycles. In the final step of the identification phase, the score fusion method combines the gait cycle recognition probability scores obtained from the trained LSTM model. Each subsystem of the proposed system is discussed as follows.

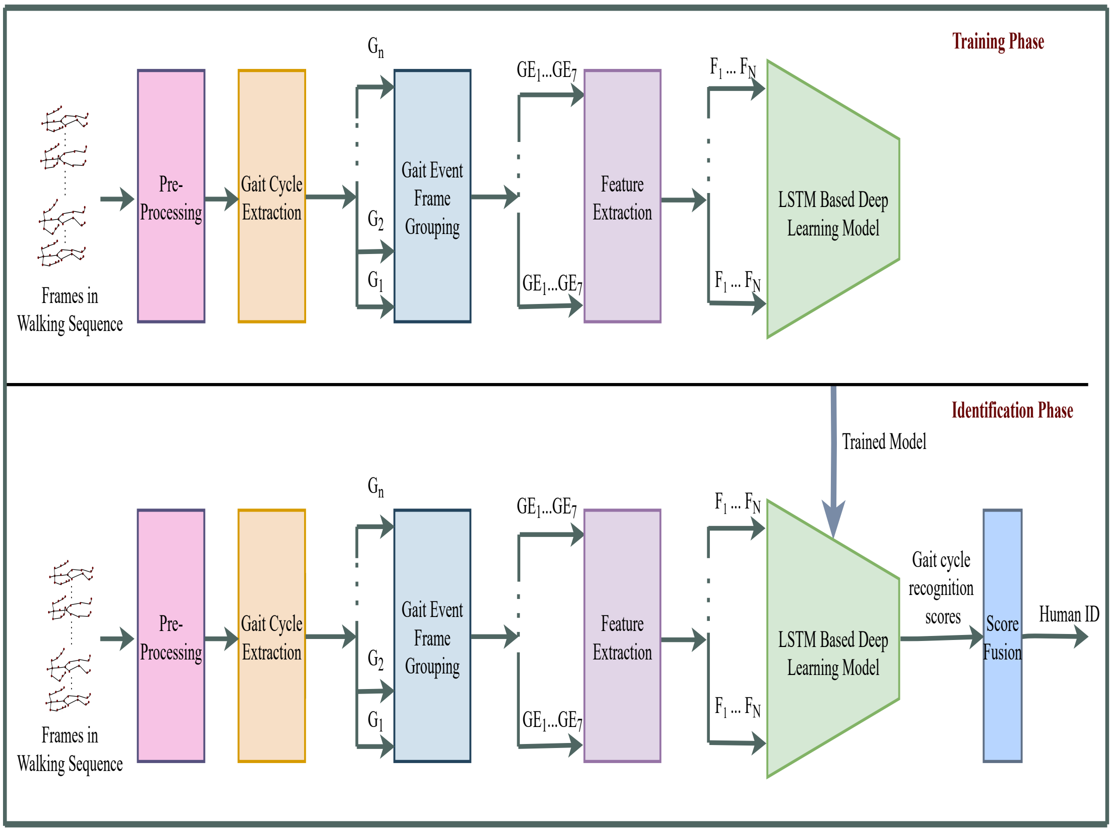


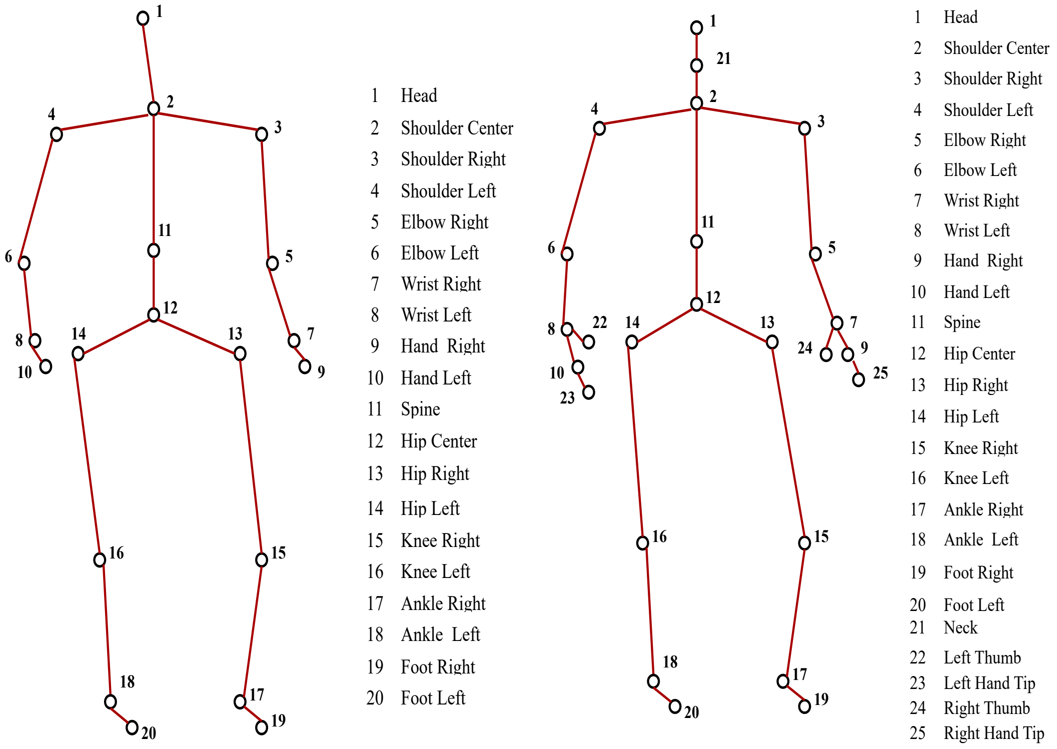
Figure 5.1: The workflow of the proposed human identification system using 3D skeleton-based gait features and the LSTM model.

5.1.1.1 Pre-processing

When a still camera is used to record the walking sequence, the distance between the subject and the camera shifts as the walk progresses. The Kinect depth sensor generates three-dimensional skeleton data at various scales, depending on the position of the subjects within the data collection environment. Therefore, in the beginning, a few steps are taken to align the skeletons in a new global coordinate system. The specifics of this process are as follows.

Align Skeleton Data in New Global Coordinate System

Consider a sequence of skeleton frames, $F = \{f_1, f_2, ..., f_n\}$ where, n is the number of frames in the entire walking sequence. Each frame contains information about a set of P 3D joint coordinates of the skeleton, $J = \{j_1, j_2..., j_P\}$. Each joint includes the three coordinates (x, y, z). Figure 5.2 depicts the skeleton joints captured by the Kinect depth sensor v1 and v2. As the person moves, these coordinate values of skeleton joints



(a) (b)

Figure 5.2: Skeleton joints captured by Kinect depth sensor. (a) 20 skeleton joints by Kinect v1. (b) 25 skeleton joints by v2.

will change. Initially, a new global coordinate system G_c is created by making the hip-center the origin and fixed at $(0, 0, 0)$. The subsequent steps translate all other joints to the new global coordinate system.

Let (h_x, h_y, h_z) be the hip center joint coordinate values along x, y, and z axes, respectively, and represented as a column vector: $\begin{bmatrix} h_x \\ h_y \\ h_z \end{bmatrix}$. To translate this joint as the origin $\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ of the global coordinate system, the translation distances are defined as $t_x = -h_x$, $t_y = -h_y$ and $t_z = -h_z$. Then all joints j_i for all $i = 1$ to P, in the original 3D coordinate system, are transformed to the new global coordinate system using the transformation matrix as shown in Equation (5.1).

$$\begin{bmatrix} x_i^{'} \\ y_i^{'} \\ z_i^{'} \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & t_x \\ 0 & 1 & 0 & t_y \\ 0 & 0 & 1 & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ z_i \\ 1 \end{bmatrix}, \tag{5.1}$$

Where, (x_i, y_i, z_i) are the initial 3D coordinate values of i^{ih} joint and $(x_i^{'}, y_i^{'}, z_i^{'})$ are its new coordinate values in G_c.

Data Normalization

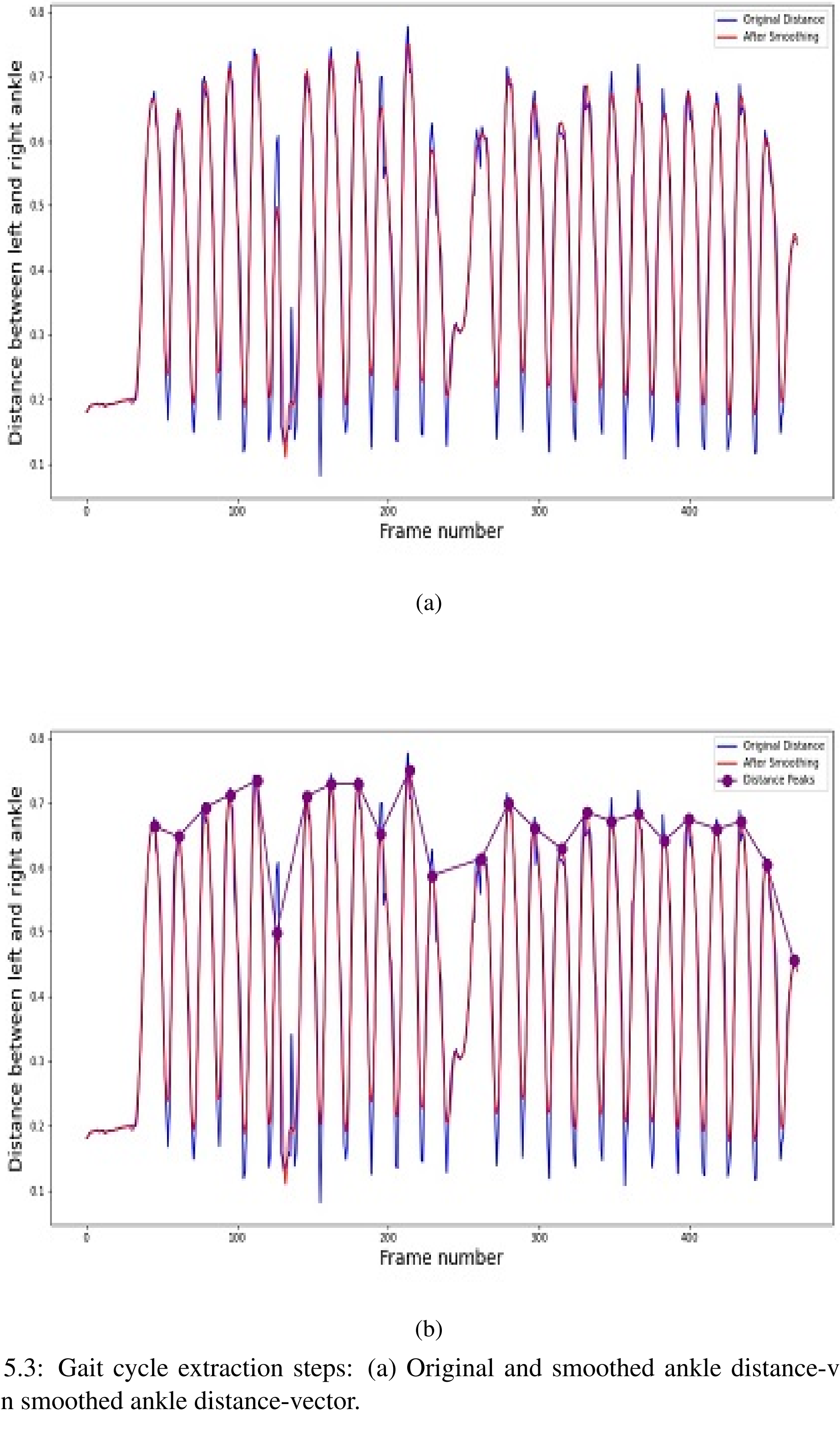
The skeleton data is normalized to reduce the impact of the variations in coordinate values. Each skeleton joint $j_i = (x_i, y_i, z_i)$ from a skeleton frame S_f is normalized using Equation (5.2).

$$j_i^{'} = \frac{j_i - c_{min}}{c_{max} - c_{min}} * 10, \forall j_i \in G_c \tag{5.2}$$

Where, $j_i^{'}$ is the normalized skeleton joint. The parameters c_{min}, and c_{max} denote the minimum and maximum coordinate values, respectively, that are computed with respect to the coordinate values in entire dataset.

5.1.1.2 Gait Cycle Extraction

Walking consists of a series of repetitive limb movements that propel the body forward. Each gait cycle begins with the IC 'Heel Strike' of one leg and concludes with the same leg's TSW 'Heel Strike'. The 'Heel Strike' is the moment in the gait cycle when the foot heel makes initial contact with the ground. So, the gait cycles can be considered by taking any leg as the reference point. During 'Heel Strike', the person's left and right ankles will be the farthest apart. The three consecutive 'Heel Strikes' make a complete gait cycle in which the first and third 'Heel Strike' are of the same leg, and the second 'Heel Strike' is of another leg. The first 60% of the gait cycle is the Stance phase, while the remaining 40% is the Swing phase. Further, researchers have identified seven different events in the gait cycle, as shown in Figure 1.4. In each of these events, limbs are positioned in specific ways to exhibit characteristics that aid in gait recognition.



(a)

(b)

Figure 5.3: Gait cycle extraction steps: (a) Original and smoothed ankle distance-vector (b) Peaks in smoothed ankle distance-vector.

In the proposed work, the Euclidean distance between the left and right ankles in each frame is calculated, and the distance vector A_D is used to record the peaks. The proposed research investigates the gait cycle starting with both legs. Further, the distance vector A_D is smoothened using Savgol filter (Savitzky and Golay, 1964). To determine gait cycles, only local peaks with a height greater than $T \times max\_peak$ are considered. Where, T is a dataset-dependent threshold. The set of frames that do not contribute to any gait cycle are discarded based on this distance vector. Figure 5.3 depicts the plots of ankle distances versus frame number, as well as the ankle distance peaks detected. The blue lines represent the initial curves. The smoothed curves are represented by red curves. Where max_{peak} represents the maximum distance between the ankles at the peak of the walking sequence.

5.1.1.3 Gait Event Frame Grouping

During different phases of the gait cycle, the human gait exhibits distinct limb movements. The authors of (Choi et al., 2019) and (Khamsemanan et al., 2018) demonstrated that each gait cycle frame contributes to the gait feature vector. The gait cycle events are also distinct for each person. Therefore, the proposed system for HI treats gait events as timestamps and extracts gait event-specific features. Figure 1.4 illustrates the timing distribution for each gait event in the entire gait cycle. The initial distribution of frames comprising a gait cycle is based on the duration of the gait event. Next, feature vectors of equal size are extracted from each group. The proposed work calculates multiple sets of inter and intra-frame joint distance and angle features for each group. To reduce the impact of occluded joints in the frame, we used the mean and standard deviation of corresponding features collected within the frame group rather than the raw features collected from the frames. The following subsections discuss the feature extraction.

5.1.1.4 Feature Extraction

During walking, each individual exhibits distinctive upper and lower limb movements. As limbs move, the distance and angle between the 3D joint coordinates of the skeleton change. The proposed work extracts a set of inter and intra-frame Euclidean distances and angles between three-dimensional joint coordinates. To reduce the impact of occluded joints in the frame, we used the mean and standard deviation of corresponding features collected within the frame group rather than the raw features collected from the frames. The Euclidean distance between any two 3D joints and angle $\Theta = \angle ABC$ where A, B, and C are three 3D joints are computed using Equations 4.1 and 4.2, respectively.

Body Part Length Features (BpLF)

In the proposed work, seven body part lengths were considered as features. The lengths of the various body parts considered are depicted in Figure 5.4 (a). The variance in these lengths must be minimal in a series of frames. So, initially, the body part length is computed for each frame fi in a gait event, and then the mean of each of these body part lengths from frames in the gait event is computed.

Joint Distance Features (JDF)

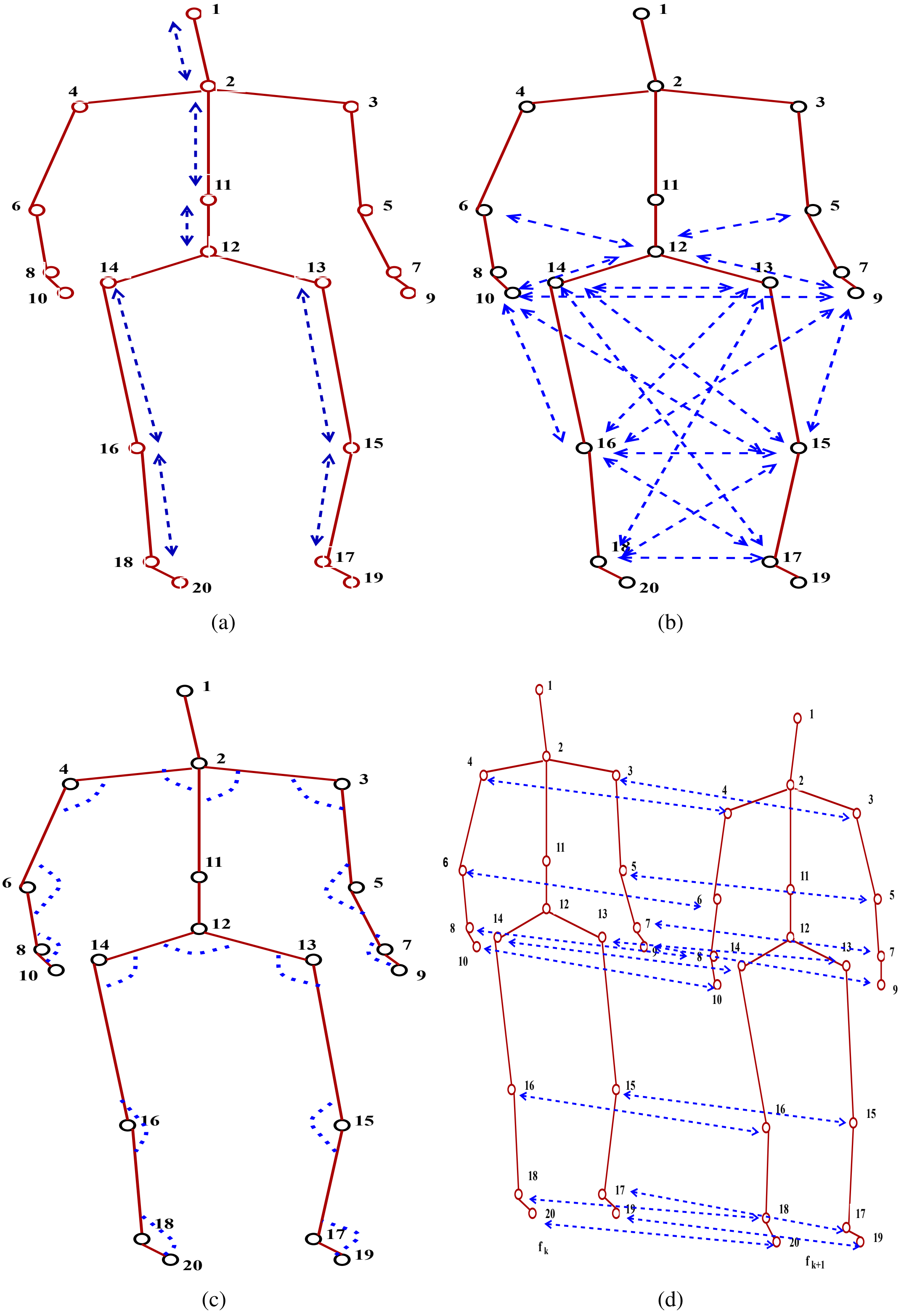
As a person walks, the relative distance between body joints dramatically changes. Consequently, the mean and standard deviation of body joint distances from frames in a gait event are calculated. Figure 5.4 (b) depicts the joint distances used to feature construction. Altogether 18 mean and 18 standard deviation features are extracted from each gait event.

Joint Angle Features (JAF)

In the proposed work, focus on the 15 joint angles shown in Figure 5.4 (c) to better understand how the body's joints interact while a person walks. As a part of analyzing gait events, we calculate the mean and standard deviation of the joint angle i from frames in each gait event. As a result, this adds 30 new features to the feature pool.

Inter-Frame Joint Distance Features (InJDF)

These are proposed to monitor the relative change in a joint position between successive frames. The Inter-Frame Joint Distance Features (InJDF) is calculated using the Equation (5.3). Where, InD is the inter-frame joint distance of skeleton joint i between the k^{th} and $(k+1)^{th}$ frames in a gait event, and $k = 1, 2, ..., n-1$, where, n is the number of frames in the gait event. Finally, these distances' mean and standard deviation are computed for each joint i considered during each gait event. Figure 5.4 (d) illustrates the sixteen inter-frame joint distances considered in this study. This adds 32 characteristics to the feature vector.



(a) (b)

(c) (d)

Figure 5.4: Extracted features. (a) BpLF. (b) JDF. (c) JAF. (d) InJDF.

$$InD = \sqrt{(x_i^k - x_i^{k+1})^2 + (y_i^k - y_i^{k+1})^2 + (z_i^k - z_i^{k+1})^2}, \tag{5.3}$$

Inter-Frame Joint Angle Features (InJAF)

To track the change in joint angle between consecutive frames, the Inter-Frame Joint Angle Features (InJAF) features are proposed. To begin, in each frame, the joint angles shown in Figure 5.4 (c) are calculated. After that, we consider subsequent frames to calculate the angular disparity. Equation (5.4) describes the entire process. Where A, B, and C are 3D joints in the skeleton, k is the number of frames in the gait event, and InA_B is the difference between $\angle ABC$ in the k^{th} and $(k+1)^{th}$ frames. Ultimately, we calculate the average and standard deviation of the angles that differ during the gait event. It adds 30 features to the feature vector.

$$InA_B = cos^{-1}\left(\frac{\vec{A^{k+1}B^{k+1}}.\vec{C^{k+1}B^{k+1}}}{\|\vec{A^{k+1}B^{k+1}}\|\|\vec{C^{k+1}B^{k+1}}\|}\right) - cos^{-1}\left(\frac{\vec{A^{k}B^{k}}.\vec{C^{k}B^{k}}}{\|\vec{A^{k}B^{k}}\|\|\vec{C^{k}B^{k}}\|}\right), \tag{5.4}$$

5.1.1.5 LSTM Based Deep Learning Model

Gait events occur in a specific order in a gait cycle. We proposed a DL model with two layers of LSTM to efficiently learn the temporal features of a gait cycle for gait recognition. Figure 5.5 depicts the proposed LSTM-based DL model. The proposed LSTM-based DL model is a sequential model implemented in Keras (Chollet, 2015). The full model consists of two LSTM layers, a Dense layer with N units and $tanh$ activation function, two Batch Normalization (BN) layers, and a Softmax layer. Since each gait cycle is divided into seven discrete events, the input data is structured accordingly. One-hot encoding is used to feed the neural network the labels used for HI. Features from gait cycles were reshaped into $(7, N)$, where N is the number of features from each gait cycle event before being provided to the first LSTM layer. The first BN layer is applied to the output of the first LSTM layer to control the overfitting of the network. The normalized output of the first BN layer is processed by the second LSTM layer. In addition, the output of the second LSTM layer is processed by the Dense layer, which in turn improves recognition performance. With the second BN layer placed before the

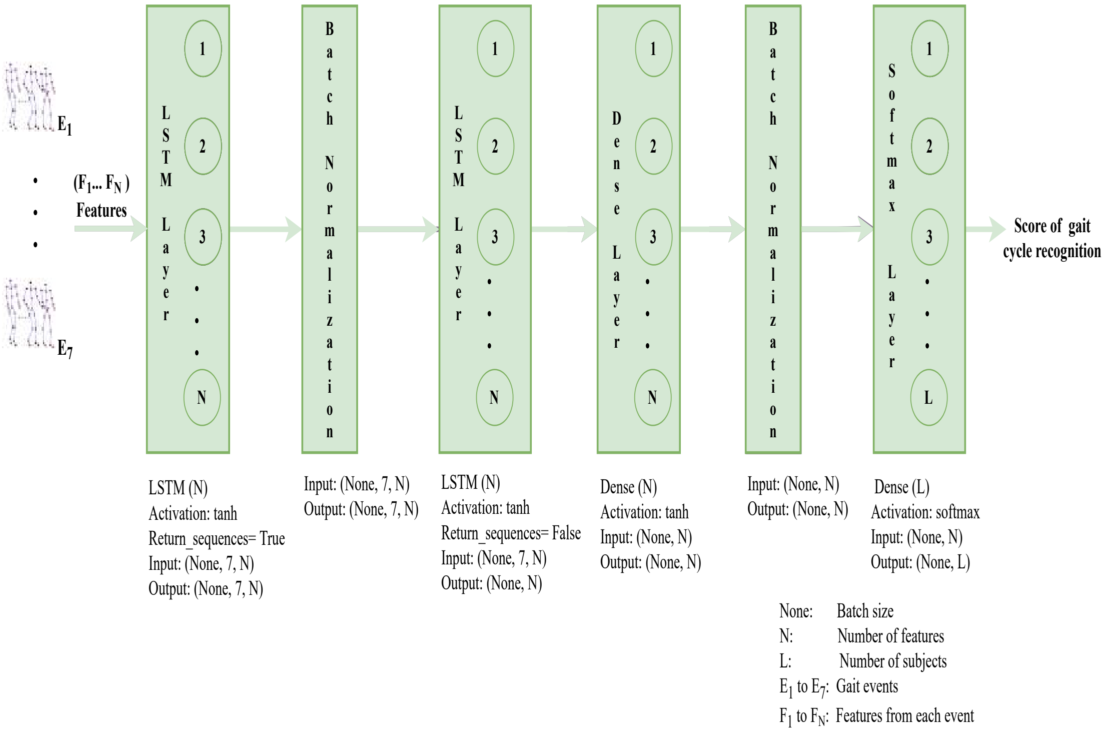


Figure 5.5: Proposed LSTM based deep learning model.

Softmax layer, training time is reduced by standardizing the output of the Dense layer. A probability score from the Softmax layer then provides the classification score of gait cycles. The optimum weights are calculated using the Adam optimizer (Kingma and Ba, 2014).

5.1.1.6 Long Short-Term Memory and Gated Recurrent Unit

LSTM (Hochreiter and Schmidhuber, 1997) is effective in combating the vanishing gradient problem of RNN and has been demonstrated its potency in learning temporal features. The schematic diagram of LSTM cell is depicted in Figure 5.6. Input gate I_t, forget gate F_t, output gate O_t, hidden state h_t, and memory cell state c_t are the various vectors that define the LSTM at each time step t. Equations (5.5) through (5.9) define each of these vectors.

$$I_t = \sigma(W_{xi}X_t + W_{hi}h_{t-1} + b_i) \tag{5.5}$$

$$F_t = \sigma(W_{xf}X_t + W_{hf}h_{t-1} + b_f) \tag{5.6}$$

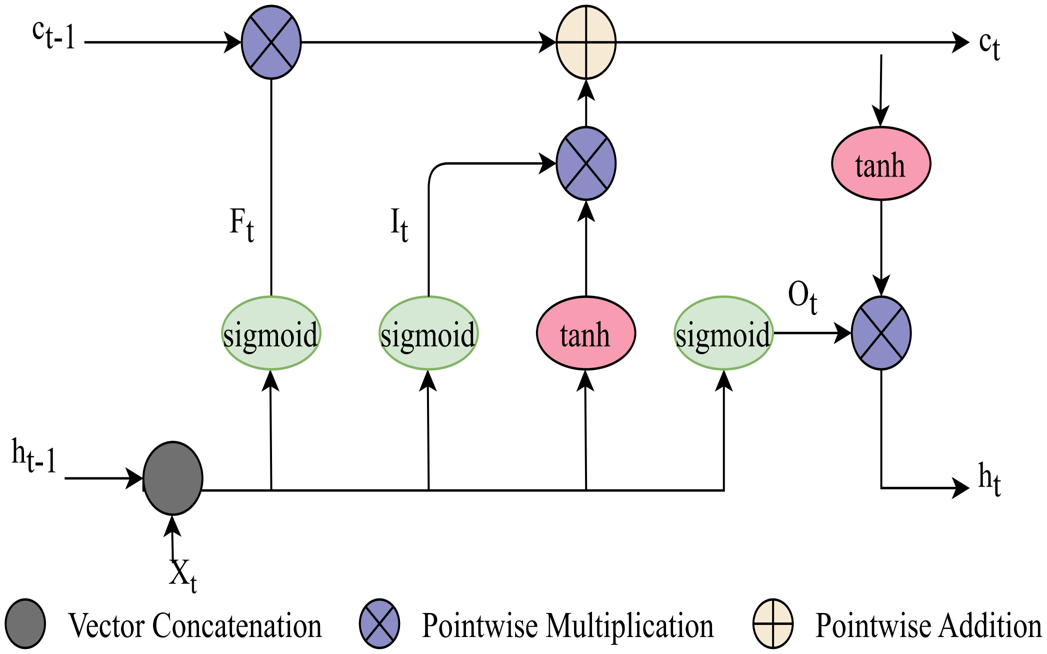


Figure 5.6: Structure of LSTM unit.

$$O_t = \sigma(W_{xo}X_t + W_{ho}h_{t-1} + b_o) \tag{5.7}$$

$$c_t = F_t \odot c_{t-1} + I_t \odot tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{5.8}$$

$$h_t = O_t \odot tanh(c_t) \tag{5.9}$$

Where, the F, I, and O determine which information will be discarded, collected, or outputted. Also, W_{xi}, W_{xf}, and W_{xo} represent the weight matrices that are updated during training. In addition b_i, b_f, and b_o represent the bias.

(Cho et al., 2014) created GRU as a gating mechanism for RNN. Comparable to LSTM, the GRU lacks an output gate and thus has fewer parameters. GRU uses 'update gate' and 'reset gate' to solve the standard RNN problem of vanishing gradients. These two vectors essentially determine which data is passed to the output. The structure of a GRU unit is depicted in Figure 5.7. The vectors in GRU are mathematically defined as follows from Equations (5.10) through (5.11).

$$z_t = \sigma(W_{xz}X_t + W_{hz}h_{t-1} + b_z) \tag{5.10}$$

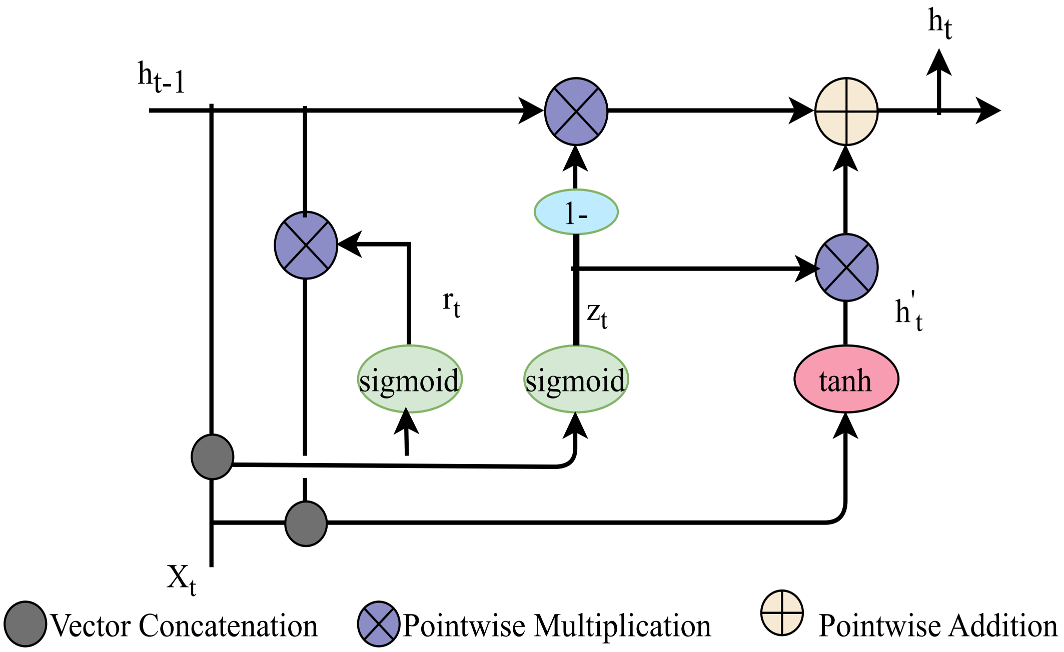


Figure 5.7: Structure of GRU unit.

$$r_t = \sigma(W_{xr}X_t + W_{hr}h_{t-1} + b_r) \tag{5.11}$$

$$h_t^{'} = \tanh(W_{xh}X_t + r_t \odot W_{hh}h_{t-1} + b_h) \tag{5.12}$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot h_t^{'} \tag{5.13}$$

Where, z_t, r_t, $h_t^{'}$, h_t, W and b represent the update and reset gate, a candidate activation vector, output vector, weights, and bias, respectively.

5.1.1.7 Score Fusion

The final Softmax layer in the proposed DL model computes the probability score vector $P = \{p_1, p_2, ..., p_L\}$ for gait cycle classification. Combining these scores, as shown in Equation (5.14), yields the final identification prediction considering gait cycles in the entire walking sequence.

$$P_{id} = argmax(\sum_{i=1}^{g} p_{i1}, \sum_{i=1}^{g} p_{i2} ..., \sum_{i=1}^{g} p_{iL}), \tag{5.14}$$

Where, g, L, p_i, and P_{id} are the number of gait cycles extracted from the walking sequence, the number of subjects in the dataset, a score of identification in i^{th} gait cycle for a person, and the person identified from the walking sequence, respectively. The function $argmax$ gives the position of the maximum score. The overall steps in the proposed system is described in Algorithm 5.1.

Algorithm 5.1: Human identification from skeleton data based gait dataset

```
Function Human_identification ($W$);
Input  : A set of walking sequences of people $W$
Output: Person identified $P_{id}$
foreach walking sequence $w \in W$ do
    Extract sequence of skeleton frames $F$
    $N$= Number of frames in $F$
    Align skeleton data from each frame in a global coordinate system using
      Equation (5.1).
    Normalize the skeleton data from each frame using Equation (5.2).
    Extract a set of gait cycles $G$ from $F$ based on distance between ankles as
      given in Section 5.1.1.2 (Page No. 78).
    foreach gait cycle $g \in G$ do
        Group frames in $g$ under different gait events:
          $E = \{LR, MST, TST, PSW, ISW, MSW, TSW\}$ based on
          proportion of frame distribution as mentioned in Table 5.1.
        Assign Gait cycle feature vector $GFE = []$
        foreach gait event $e \in E$ do
            Assign gait event feature vector $FE = []$
            //Perform feature extraction
            Extract $BpLF$, $JDF$, $JAF$, $InJDF$, and $InJAF$ features using
              the procedure given in Section 5.1.1.4 (Page No. 80)
            $FE = [BpLF, JDF, JAF, InJDF, InJAF]$
            $GFE = [GFE, FE]$
Partition the feature vectors into train and test sets based on specifications of
  Evaluation Protocol.
Train the proposed LSTM based DL model.
Feed feature vector of test set to fitted proposed LSTM based DL model.
if Score fusion required based on Evaluation Protocol then
    Perform Score fusion as given in Equation (5.14).
    Return Person Identified $P_{id}$.
else
    foreach gait cycle $g \in test$ do
        Return Person identification label which obtained highest score.
```

5.1.2 Experiments, Results, and Analysis

To investigate how various features contribute to gait recognition, the LSTM model is trained with various feature combinations during the training phase. The fitted LSTM model is utilized during the identification stage for gait cycle recognition. Next, based on its suitability for evaluation protocols, the proposed score fusion is applied to the Softmax layer's output. In some evaluation protocols, score fusion is omitted because it serves no purpose. A series of experiments are conducted on all the datasets using various classification algorithms, including KNN, SVM, RF, two-layer ANN, and GRU (Cho et al., 2014).

5.1.2.1 Datasets

The proposed system's performance is assessed using four publicly available benchmark 3D skeleton-based gait datasets, namely: Kinect Gait Biometry Dataset (KGBD) (Andersson and Araujo, 2015), University of Patras Computer Vision-1 (UPCV1) (Kastaniotis et al., 2015; Kastaniotis et al., 2013), University of Patras Computer Vision-2 (UPCV2) (Kastaniotis et al., 2016), and VisLab Multi-View Kinect Skeleton (KS20) dataset (Nambiar et al., 2017a,b).

Kinect Gait Biometry Dataset (KGBD)

The KGBD is a compilation of the 3D skeleton-based gait data from 164 subjects. Kinect v1 was positioned in the center of a semicircle with a spinning dish to track the subject's movement. The subjects were directed to walk in a semicircular path during data collection. The majority of walking sequences are comprised of 500 to 650 frames. Each participant, with a few exceptions, has five walking sequences. Each frame includes the 3D coordinates of 20 joints. The number of gait cycles extracted per walking sequence is between 8-29.

UPCV1

Using the Kinect v1 depth sensor, the UPCV1 gait dataset is collected. It includes five walking sequences from 15 male and 15 female participants. Each person walks in a straight line. The number of frames in the walking sequence ranges from 50 to 120. Most walking sequences are comprised of two gait cycles.

UPCV2

The UPCV2 gait dataset was collected using a Microsoft Kinect v2 depth sensor, and each skeleton has 25 joints. It contains ten recordings of 30 individuals walking in a straight line (17 males and 13 females). The majority of walking sequences consist of three gait cycles.

KS20 VisLab Multi-view Kinect Skeleton Dataset

It is a dataset of gait captured from multiple views with the Microsoft Kinect v2 depth sensor. The walking sequences of 20 subjects were captured from five different views ('left-lateral'→0°, 'left-diagonal'→ 30°, 'frontal'→ 90°, 'right-diagonal'→ 130°, and 'right-lateral'→ 180°). Three samples are taken from each viewpoint per subject, with each sample containing a single gait cycle.

5.1.2.2 Evaluation Protocols

Single-view Dataset

The performance of proposed system on three single-view datasets KGBD, UPCV1, and UPCV2 is evaluated using the following two benchmark evaluation protocols.

- $Protocol - 1$ $(Leave - one - sequence - out)$: Human identification using the walking sequence based on the approach of (Khamsemanan et al., 2018). The dataset is divided into K distinct folds based on the number of walking sequences for each individual. Therefore, 5-fold cross-validation is employed on KGBD and UPCV1 datasets and 10-fold cross-validation on the UPCV2 dataset. Each person's walking sequence is distributed across multiple folds so that each fold contains gait cycles of one walking sequence. Then, K models are constructed iteratively using the i^{th} fold as the testing fold in the i^{th} iteration.

- $Protocol - 2$ $(Five - fold - cross - validation)$: The protocol described in (Bari and Gavrilova, 2019) is used to assess the performance of the proposed system in recognizing the gait cycle. As the score fusion subsystem plays no role in gait cycle recognition, it is omitted in the experiments. Again, we employed the K-fold cross-validation method, with K equal to 5. Contrary to $Protocol - 1$, where all gait cycles extracted from a walking sequence fall into the same fold, in $Protocol - 2$, the gait cycles extracted from a walking sequence are randomly distributed across all five folds. Five models are iteratively created with i^{th} fold for testing in the i^{th} iteration.

Multi-view Dataset

The proposed system is tested on a multi-view gait dataset using the following two evaluation protocols.

- $Random-split$: Randomly selected two samples from each viewpoint are used for training, and the remaining one from each viewpoint is used for testing.
- $Cross-view-split$: Test on all samples from each viewpoint while considering samples from remaining viewpoints for training.

5.1.2.3 Evaluation Metrics

The effectiveness of the proposed work is evaluated using four metrics: $Accuracy$, $Precision$, $Recall$, and $F-Score$. The $Precision$ reflects the proportion of accurately predicted positives. $Recall$ quantifies a model's capability to predict positive outcomes. The $F-Score$ quantifies the harmonic mean of $Precision$ and $Recall$. All these metrics are defined using the Equations (5.15) to (5.18). Where TP, TN, FP, and FN represent $True\,Positive$, $True\,Negative$, $False\,Positive$, and $False\,Negative$, respectively. Additionally, the capability of the proposed system at various Rank levels is evaluated using the CMC. CMC is a plot of cumulative recognition performance across various Rank levels.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \tag{5.15}$$

$$Precision = TP/(TP + FP) \tag{5.16}$$

$$Recall = TP/(TP + FN) \tag{5.17}$$

$$F - Score = \frac{2 * Precision * Recall}{(Precision + Recall)} \tag{5.18}$$

5.1.2.4 Training Parameters

Each model is trained for 150 epochs with a batch size of 64 and an initial learning rate of 0.001. The learning rate is decreased by 0.001 based on observed validation loss, resulting in a more stable model. The model is trained with Categorical Cross-Entropy loss function, and Adam optimizer (Kingma and Ba, 2014).

5.1.2.5 Experiments on Single-view Dataset

This section elaborates on the observed results in various experiments conducted on single-view datasets using $Leave-one-sequence-out$ $(Protocol-1)$ and $Five-fold-cross-validation$ $(Protocol-2)$ evaluation protocols.

Leave-one-sequence-out

Table 5.2 demonstrates the Rank-1 performance of the HI system employing the entire walking sequence with various feature combinations and the proposed LSTM-based DL model on three benchmark datasets. The LSTM model is created with four different combinations of features. The number of gait event features utilized indicated between brackets. The combination of body part length and additional features yielded an $Accuracy$ of greater than 95% across all three datasets. Additionally, the GRU-based DL model with 105 features performed admirably on all datasets.

Table 5.2: Rank-1 recognition performance (in %) of the proposed HI system using $leave-one-sequence-out$ in various experimental setups (Best results are in bold).

Experimental Setup	**Gait Dataset**											
	KGBD				**UPCV1**				**UPCV2**			
	Accuracy	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
InJDF (30 Features) + LSTM	72.9	63.2	73	65.8	59.99	49	60	52.4	80.32	72.3	80.5	74.7
InJDF + InJAF (62 Features) + LSTM	91.32	87.4	91	88.4	59.99	48.8	60	52.4	88.32	83.3	88.5	84.8
BpLF+JDF+ JAF (73 features) + LSTM	97.56	96.6	97.6	96.66	95.9	94	96	94.8	98.33	97.5	98.4	97.9
BpLF+JDF+ InJDF+JAF (105 Features)) + GRU	98.53	97.8	98.6	98	95.99	94	96	94.8	97.33	96	97.4	96.6
BpLF+JDF+ InJDF+JAF (105 Features)) + LSTM (Proposed)	**98.54**	**97.6**	**98.6**	**97.8**	**98.00**	**97**	**98**	**97.4**	**98.33**	**97.5**	**98.5**	**98**

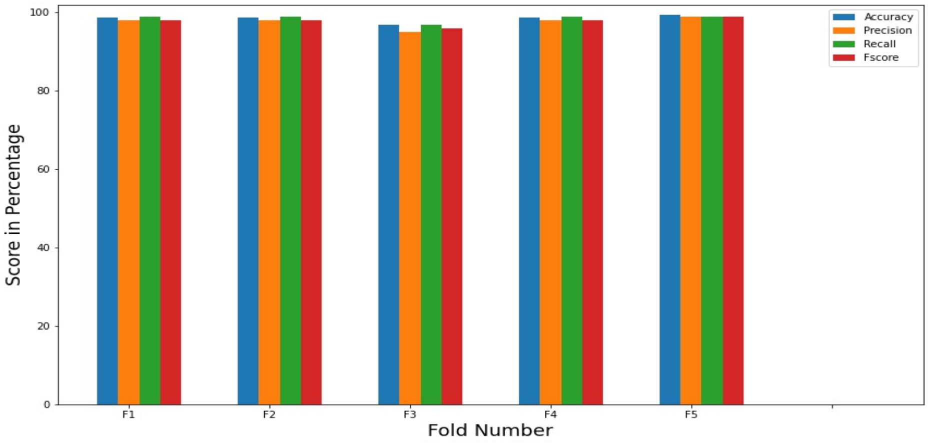
Without the Body-part Length Features (BpLF) in the feature combination, the performance of the proposed system is drastically diminished. Using 105 features, the LSTM model achieved 98.54%, 98.00%, and 98.33% *Accuracy* in the KGBD, UPCV1, and UPCV2 datasets, respectively. Moreover, the performance of the model trained with 73 features across all datasets is promising. However, the LSTM model with 105 features consistently recorded the highest value for all evaluation metrics. The LSTM models with 62 and 30 features did not achieve satisfactory performance. It is observed that the 105-feature GRU model almost matched the 105-feature LSTM model's performance on KGBD but fell short on the other two datasets. Figure 5.8 illustrates the proposed system's performance in all K folds of the dataset. Despite a small variance in values, the model performed nearly identically across all test folds.

Cumulative Match Characteristic (CMC) Curve

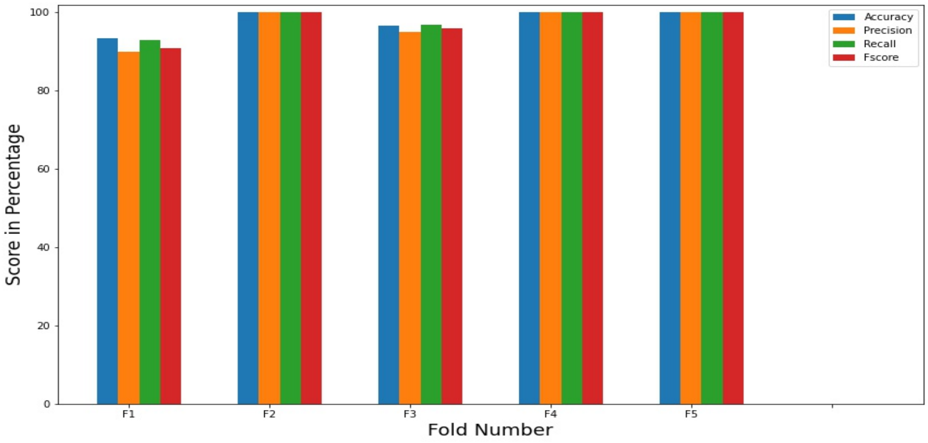
Before identifying the criminal, investigation agencies must consider a group of suspicious people. The CMC test is performed to assess the proposed system's performance in different Ranks. Figure 5.9 depicts the CMC curves of the proposed system for three datasets. It compares the *Accuracy* of LSTM and GRU-based DL models at different Ranks. The proposed system with LSTM demonstrated 98.53%, 98.00% , and 98.33% Rank-1 *Accuracy* on KGBD, UPCV1, and UPCV2, respectively. Further, on KGBD, at Rank-2 itself, it reached 99.27% *Accuracy*. On UPCV1, the Rank-3 recognition *Accuracy* is 100%. On UPCV2, the proposed system achieved the recognition *Accuracy* 99.67% at Rank-3. Overall, the proposed system achieved recognition rates of over 99% within the top three Ranks, making it suitable for real-world applications.

Five-fold cross-validation

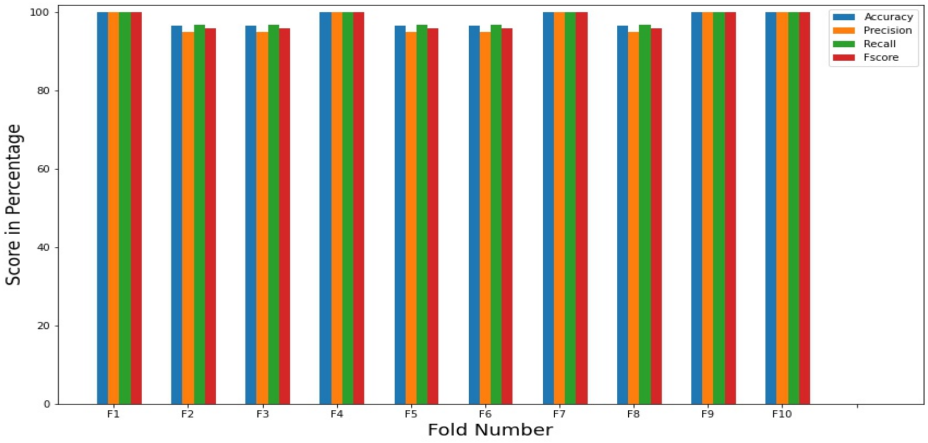
The effectiveness of the proposed system at recognizing the gait cycle is determined through a $Five-fold-cross-validation$ evaluation. Due to its irrelevance, score fusion has been omitted in this. Table 5.3 demonstrates the performance of Rank-1 gait cycle recognition *Accuracy* in various experimental setups utilizing $Five-fold-cross-validation$ on the KGBD, UPCV1, and UPCV2 datasets. The proposed DL model with 105 features performs the best in all three datasets. The 105-feature LSTM-based DL model achieved 97.12%, 96.30%, and 99.46% *Accuracy* on the KGBD, UPCV1, and UPCV2 datasets, respectively. This model also scored the highest *Precision*, *Recall*,



(a)

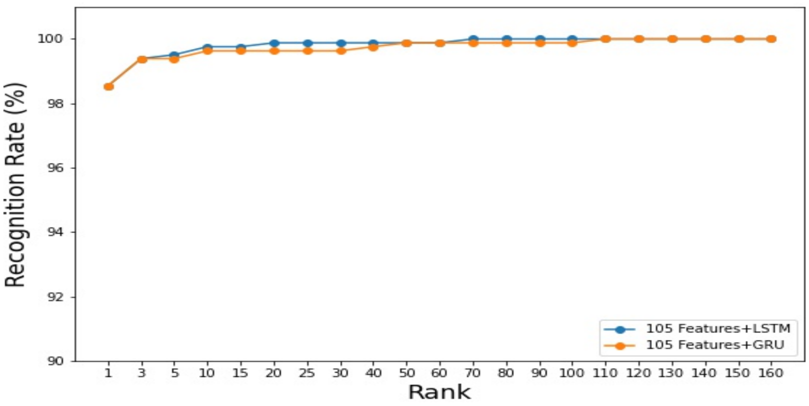


(b)

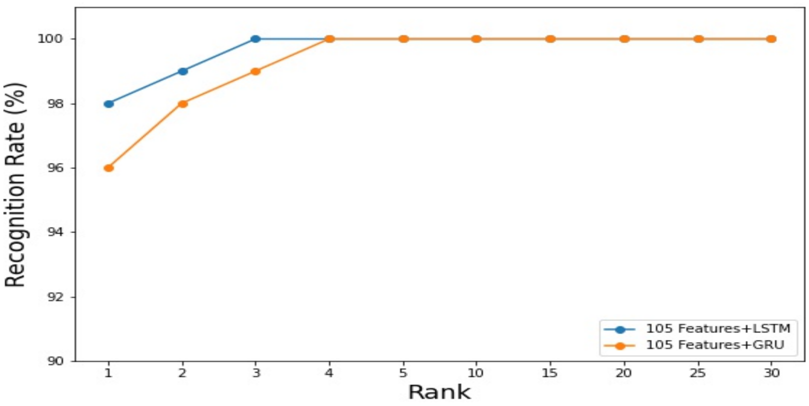


(c)

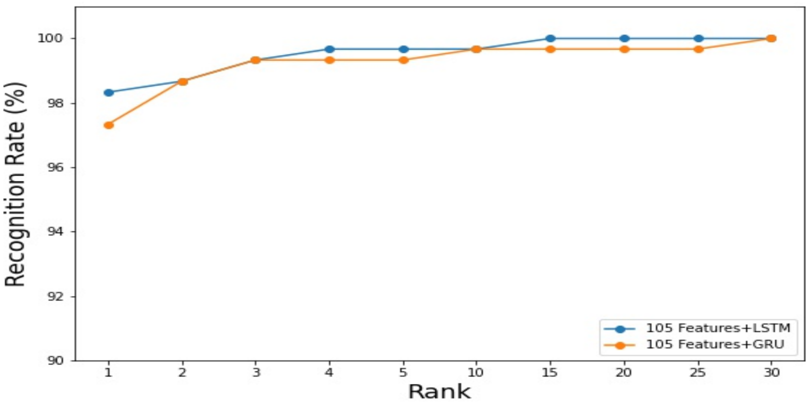
Figure 5.8: Rank-1 recognition performance of the proposed HI system using $Leave - one - sequence - out$ in K folds. (a) KGBD. (b) UPCV1. (c) UPCV2.



(a)



(b)



(c)

Figure 5.9: The CMC curves obtained using $Leave-one-sequence-out$ evaluation protocol. (a) KGBD. (b) UPCV1. (c) UPCV2.

and $F1 - Score$. Several experiments were conducted to compare the performance of these 105 features to other ML classification methods, including KNN, RF, SVM, and 2-layer ANN model. None of these techniques surpassed the performance of the proposed LSTM model with 105 feature combinations. In addition, the performance of the 2-layer ANN model was significantly inferior to that of the proposed LSTM model. On the UPCV2 dataset, the GRU-based DL model performed similarly to the LSTM model, whereas it performed marginally worse on the KGBD and UPCV1 datasets. Thus, the proposed system demonstrated the classification efficacy of the proposed features in combination with LSTM-based DL model. In addition, the LSTM model with all 135 features performed worse than the LSTM model with 105 features on KGBD and UPCV1. However, this performance is nearly identical on UPCV2. Also, the 135 features will add more complexity to the DL model than 105 features. Thus, it has been demonstrated that the proposed LSTM model trained with 105 features demonstrates superior performance in recognizing gait cycles.

Cumulative Match Characteristic (CMC) Curve

Figure 5.10 depicts the CMC test curves of the proposed system utilizing $Five-fold-cross-validation$ evaluation protocol. Also, the $Accuracy$ of the proposed LSTM-based system is compared to the $Accuracy$ of the GRU-based system at various Ranks. The proposed system achieved a 97.12%, 99.3%, and 99.86On UPCV1 dataset, the Rank-1, Rank-3, and Rank-10 $Accuracy$ are 96.30%, 97.85%, and 99.38%, respectively. On UPCV2, the CMC test reported recognition $Accuracy$ of 99.46%, 99.68%, and 99.89% for Rank-1, Rank-2, and Rank-3, respectively. The CMC test revealed that the proposed system with 105 features achieved recognition $Accuracy$ greater than 99.9% at lower-level ranks.

Comparison with State-of-the-art Existing Works

Tables 5.4, 5.5, and 5.6 compare the Rank-1 $Accuracy$ of the proposed work using $Leave-one-sequence-out$ evaluation protocol to state-of-the-art existing works on the KGBD, UPCV1, and UPCV2 datasets, respectively. Table 5.4 report that the proposed system outperformed five existing works on KGBD. Proposed system showed 1.15%, 1.04%, 1.98%, 3.14%, and 10.84%, better $Accuracy$ than (Liu et al., 2019), (Khamsemanan et al., 2018), (Li et al., 2017), (Yang et al., 2016), and (Andersson and

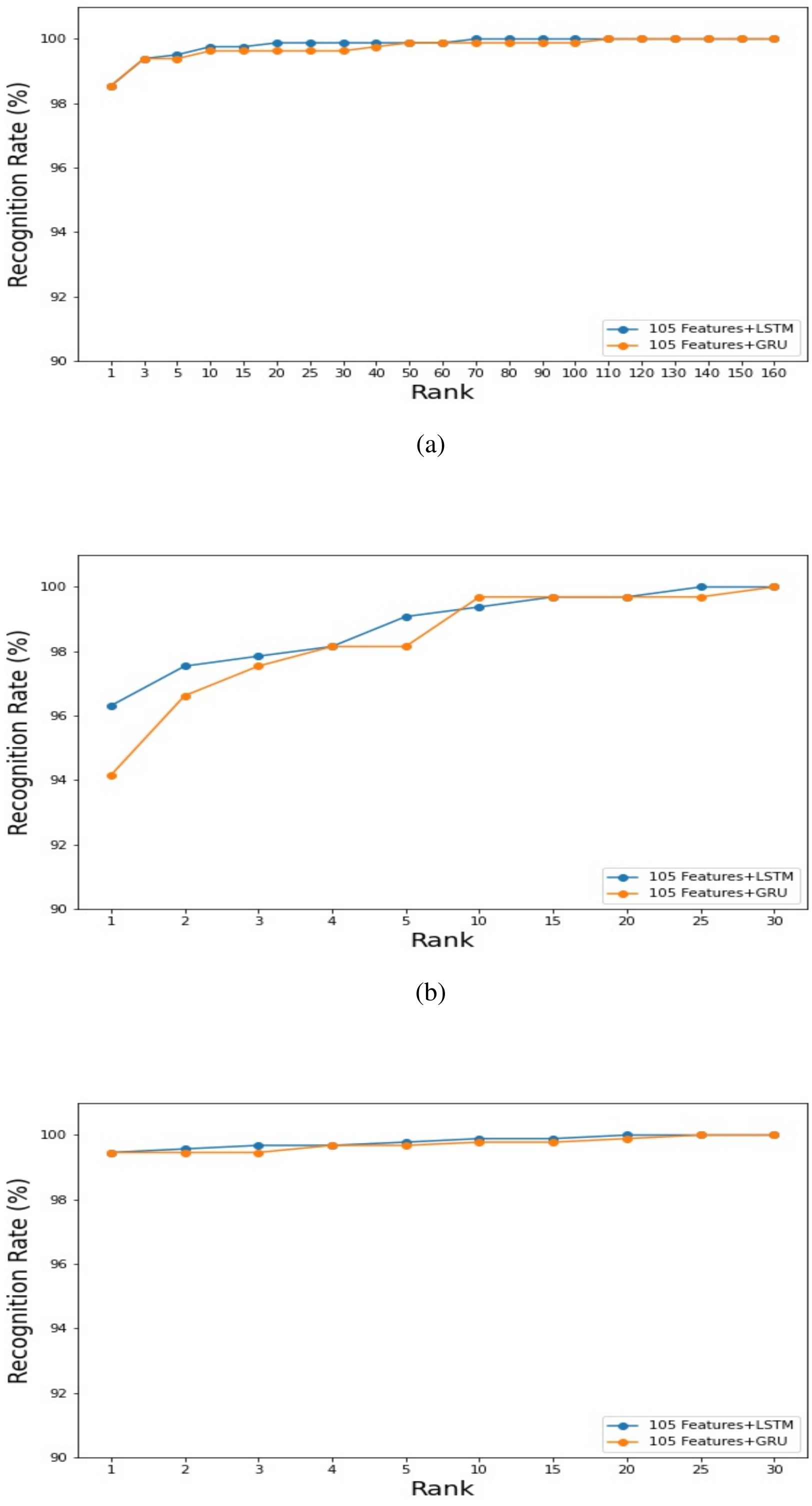
Table 5.3: Rank-1 recognition performance of the proposed HI system with different feature combinations using $Five - fold - cross - validation$ (in %) (Best results are in bold).

Experimental Setup	**Gait Dataset**											
	KGBD				**UPCV1**				**UPCV2**			
	Accuracy	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
BpLF+JDF+ InJDF+JAF (105 Features)+ RF	53.2	59	53.2	53.6	76.6	76	76.6	73.6	94.2	94.6	94.2	93.8
BpLF+JDF+ InJDF+JAF (105 Features)+ KNN	64.8	67.6	64.8	64.6	77.4	78.6	77.4	75.2	91.2	92.8	91.2	91.4
BpLF+JDF+ InJDF+JAF (105 Features)+ SVM	84	84.6	83.8	83.8	88.2	88.4	88.2	86.6	97.2	97.8	97.2	97.5
BpLF+JDF+ InJDF+JAF (105 Features)+ 2 Layer ANN	88.25	88.6	88.4	88.4	12.8	6.2	12.8	6.2	90.6	91.2	90.6	90.2
InJDF (30 Features) + LSTM	28.49	27.4	28.4	27	57.33	55.8	57.4	53.6	74.19	76.2	74.4	73.6
InJDF + InJAF (62 Features) + LSTM	49.29	49.2	49.2	49	54.76	54.6	54.6	51.2	82.25	84.8	82.2	82.2
BpLF+JDF+ JAF (73 Features) + LSTM	96.39	96.6	96.4	96.4	93.84	95.6	93.8	93.4	99.02	99.4	98.8	98.8
BpLF+JDF+ InJDF+JAF+ InJAF (135 Features) + LSTM	96.78	96.8	96.6	96.6	93.22	95.6	93.2	92.8	99.44	99.8	99.2	99.2
BpLF+JDF+ InJDF+JAF (105 Features) + GRU	96.36	96.8	96.2	96.2	94.14	95.4	94	93.6	**99.46**	**99.8**	**99.2**	**99.2**
BpLF+JDF+ InJDF+JAF (105 Features) + LSTM (Proposed)	**97.12**	**97**	**97**	**97**	**96.30**	**97.2**	**96.2**	**96**	**99.46**	**99.8**	**99.2**	**99.2**

Table 5.4: Comparison of Rank-1 $Accuracy$ of the proposed HI system using $Leave - one - sequence - out$ protocol with existing methods on KGBD. (Best result is in bold).

Human Identification Methods	**Accuracy (%)**
(Andersson and Araujo, 2015)	87.7
(Yang et al., 2016)	95.4
(Li et al., 2017)	96.56
(Khamsemanan et al., 2018)	97.5
(Liu et al., 2019)	97.39
Proposed System (with 105 features + LSTM)	**98.54**

Araujo, 2015), respectively. Table 5.5 reports the comparison of $Accuracy$ on UPCV1 with four existing works. On this dataset, the performance of the proposed work improved by 2.33%, 4.67%, 4.71%, and 10.2% relative to the works of (Kastaniotis et al.,



(a)

(b)

(c)

Figure 5.10: The CMC curves of the proposed system using $Five-fold-Cross-Validation$. (a) KGBD. (b) UPCV1. (c) UPCV2.

Table 5.5: Comparison of the proposed HI system's Rank-1 $Accuracy$ using $Leave - one - sequence - out$ to existing methods on UPCV1. (Best result is in bold).

Human Identification Methods	Accuracy (%)
(Ince et al., 2017)	87.8
(Kastaniotis et al., 2015)	93.29
(Rahman and Gavrilova, 2017)	93.33
(Kastaniotis et al., 2016)	95.67
Proposed System (with 105 features + LSTM)	**98.00**

Table 5.6: Comparison of the proposed HI system's Rank-1 $Accuracy$ using $Leave - one - sequence - out$ to existing methods on UPCV2. (Best result is in bold).

Human Identification Methods	Accuracy (%)
(Bobillo et al., 2017)	89.03
(Hosni and Amor, 2020)	92.41
(Kastaniotis et al., 2016)	97.05
Proposed System (with 105 features + LSTM)	**98.33**

2016), (Rahman and Gavrilova, 2017), and (Kastaniotis et al., 2015), respectively. The comparison with three existing works on the UPCV2 dataset is shown in Table 5.6, and it reports that the proposed system achieved 1.28%, 5.92%, and 9.3% higher $Accuracy$ than (Kastaniotis et al., 2016), (Hosni and Amor, 2020), andv(Bobillo et al., 2017), respectively.

The results of the proposed system using $Five-fold-cross-validation$ protocol is are compared to those found in (Bari and Gavrilova, 2019). Tables 5.7, and 5.8 report the comparison of Rank-1 $Accuracy$ on KGBD and UPCV1, respectively. On KGBD, although the proposed model's Rank-1 $Accuracy$ is slightly lower than the existing method, it reached more than 99% in Rank-2 and 99.86% in Rank-10. At the same time, (Bari and Gavrilova, 2019) achieved above 99% and 99.64% recognition $Accuracy$ in Rank-3 and Rank-10, respectively, which is slightly lower than the proposed work. In addition, the Rank-1 $Accuracy$ achieved by the proposed system on UPCV1 is superior to the existing work. According to (Bari and Gavrilova, 2019), the existing system scored more than 99% $Accuracy$ in Rank-7, whereas the proposed system scored more than 99% $Accuracy$ in Rank-5.

Table 5.7: Comparison of the proposed HI system's Rank-1 $Accuracy$ using $Five-fold-cross-validation$ to state-of-the-art-works on KGDB. (Best result is in bold).

Gait Cycle Recognition Methods	Accuracy (%)
(Bari and Gavrilova, 2019)	**98.08**
Proposed System (with 105 features + LSTM)	97.12

Table 5.8: Comparison of the proposed HI system's Rank-1 $Accuracy$ using $Five-fold-cross-validation$ to existing methods on UPCV1. (Best result is in bold).

Gait Cycle Recognition Methods	Accuracy (%)
(Bari and Gavrilova, 2019)	95.30
Proposed System (with 105 features + LSTM)	**96.30**

5.1.2.6 Experiments on Multi-view Dataset

Random-split

We performed a series of experiments on the multi-view skeleton-based gait dataset KS20 to demonstrate the effectiveness of the proposed system in view-invariant scenarios. The samples in the dataset comprise a single gait cycle. To increase the number of gait cycles per sample, data augmentation is performed. As previously explained, a gait cycle consists of three successive peaks in ankle distances. Therefore, in addition to the gait cycle provided by the data sample, an additional gait cycle beginning with a different leg is created by inserting the frames between the first and second peaks immediately after the third peak. After the translation and normalization steps outlined in Section 5.1.1.1 (Page No. 76), the number of gait cycles is increased by rotating all skeleton joints in both gait cycles about different axes. The augmented gait cycles are used only for training the DL model. Again, score fusion part was excluded during experimentation as they do not play any role in $Random-split$ and $Cross-view-split$ evaluation protocols.

With the same training parameters as the single-view datasets, a series of experiments were conducted on a multi-view dataset to understand the proposed system better. Table 5.9 displays the Rank-1 results obtained with various feature combinations on the $Random-split$ evaluation protocol. The experiment that used a combination of 105 features and LSTM to recognize people by their gait achieved the best results.

Table 5.9: Rank-1 recognition performance of the proposed HI system on KS20 using $Random - split$ protocol (in %) (Best result is in bold).

Experimental Setup	Accuracy	Precision	Recall	F1-Score
BpLF+JDF+ InJDF+JAF (105 Features)+ KNN	47	52	47	47
BpLF+JDF+ InJDF+JAF (105 Features)+ RF	48	50	48	45
BpLF+JDF+ InJDF+JAF (105 Features)+ SVM	65	68	65	64
BpLF+JDF+ InJDF+JAF (105 Features)+ 2 Layer ANN	12	2	12	3
InJDF (30 Features)+LSTM	31	29	31	29
InJDF + InJAF (62 Features)+LSTM	36.25	37	36	35
BpLF+JDF+ JAF (73 features)+LSTM	86	87	86	86
BpLF+JDF+ InJDF+JAF+ InJAF (135 Features)+LSTM	86	88	86	86
BpLF+JDF+ InJDF+JAF (105 Features)+ GRU	84	87	84	84
BpLF+JDF+ InJDF+JAF (105 Features)+LSTM (Proposed)	**88**	**89**	**88**	**88**

Cumulative Match Characteristic Curve

Figure 5.11 depicts the CMC curve obtained using $Random-split$ evaluation protocol. From this it is clear that the proposed system achieved more than 95% $Accuracy$ in Rank-5 itself.

Comparison with State-of-the-art Existing Works

Table 5.10 compares the proposed system for view-invariant gait recognition to state-of-the-art existing works employing $Random - split$ evaluation protocol. The proposed system performed better than the majority of existing methods.

Cross-view-split

Table 5.11 provides the Rank-1, Rank-5, and Rank-10 $Accuracy$ obtained using $Cross-view - split$ with 105 features and the proposed LSTM model. In addition, it reports the comparison of Rank-1 $Accuracy$ with existing methods on all views. The proposed system achieved higher recognition $Accuracy$ in two distinct views than existing methods. The results of (Liao et al., 2020) are based on published results from (Rao et al., 2021).

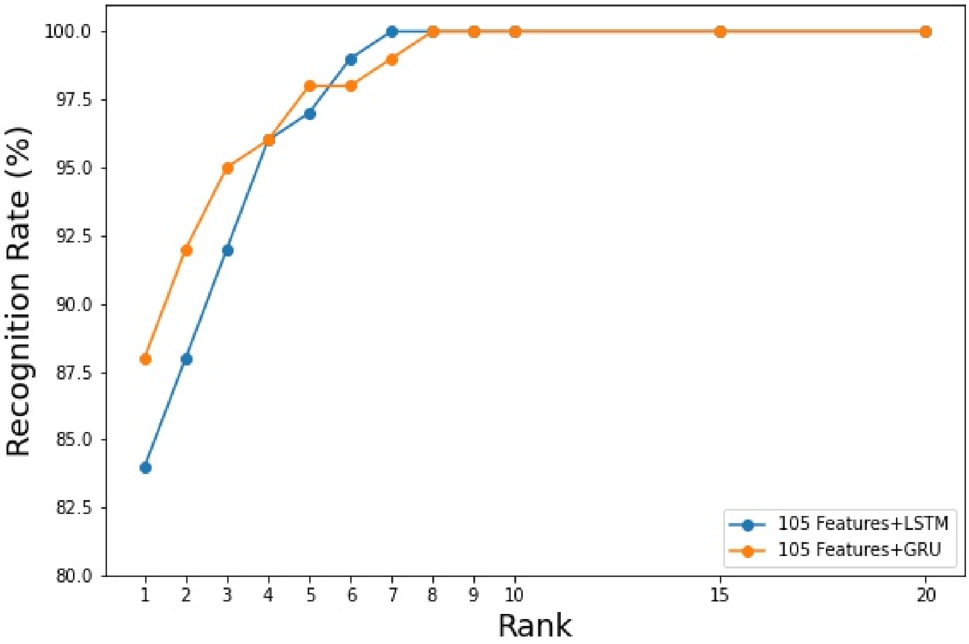


Figure 5.11: CMC test curve obtained in $Random - split$ method on KS20.

Table 5.10: Comparison of Rank-1 $Accuracy$ to state-of-the-art works on KS20 dataset using $Random - split$ protocol. (Best result is in bold).

Human Identification Method	**Accuracy (%)**
Pose Gait (Liao et al., 2020)	70.5
Context Unware (Nambiar et al., 2017a)	79.33
3D spatial-temporal (Huynh-The et al., 2020)	87.63
Context Aware (Nambiar et al., 2017a)	88.67
Self-Supervised (Rao et al., 2021)	**92**
BpLF+JDF+ InJDF+JAF (105 Features)+LSTM (Proposed)	88

5.1.2.7 Ablation Experiments

Several ablation experiments are carried out to determine the importance of layers in the proposed DL model for identifying gait cycles. The models are evaluated using $Five - fold - cross - validation$ for single-view datasets and $Random - split$ for multi-view datasets. Table 5.12 displays the results obtained by removing layers from the proposed DL model on three datasets. The results suggest that omitting any layer from the proposed model has a discernible effect on performance.

Table 5.11: $Accuracy$ (in %) of proposed HI system on KS20 dataset using $Cross - view - split$ protocol and comparison with existing methods.

View	**Pose Gait** (Liao et al., 2020)	**Self-Supervised** (Rao et al., 2021)	**Proposed System**		
	Rank-1	Rank-1	Rank-1	Rank-5	Rank-10
0°	24.6	**48.8**	25	46.67	76.67
30°	19.1	**53.6**	45	75	91.67
90°	29.7	54.9	**73.33**	93.33	98.33
130°	27.3	44.5	**78.33**	95	100
180°	25	**57.5**	36.67	78.33	90

Table 5.12: Ablation experiments results.

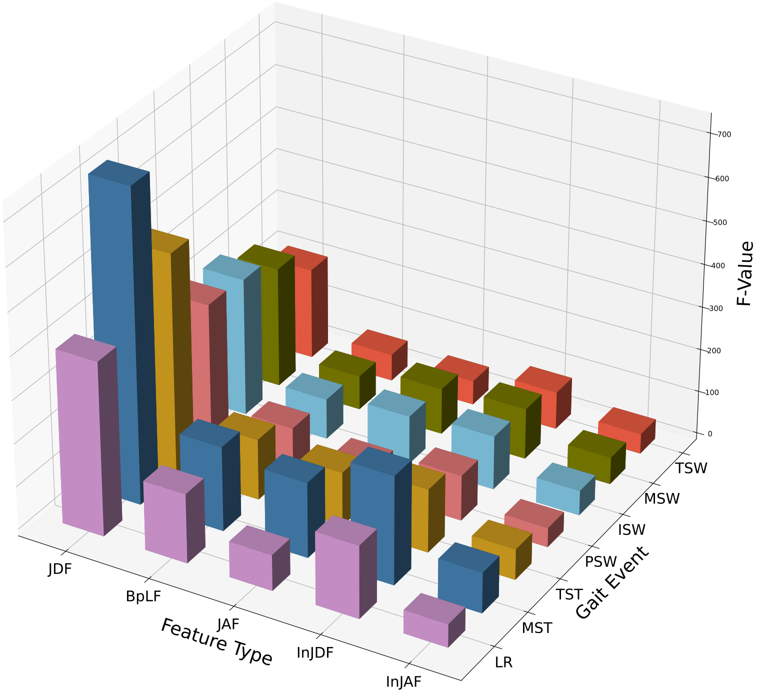
Excluded Layers	**Dataset**			
	KGDB	**UPCV1**	**UPCV2**	**KS20**
	Accuracy (%)	**Accuracy (%)**	**Accuracy (%)**	**Accuracy (%)**
Second LSTM	94.14	92	97	75
First BN	92.96	95.45	99	84
Second BN	94.86	93.85	98.82	75
First Dense	95.95	93	98.65	77

5.1.2.8 Statistical Analysis

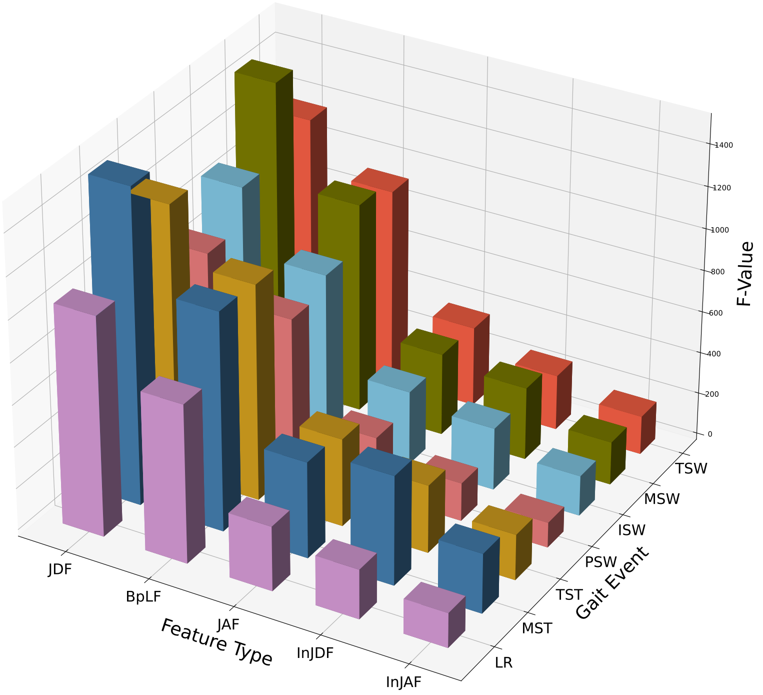
Statistical analysis is performed on features, to determine the influence of features in gait recognition. Also, analysis is done on results obtained from different folds using various combinations of features and the proposed DL model to investigate the features chosen and the methodology employed.

Feature Set Analysis

According to Analysis of Variance (ANOVA) theory, a large F-value indicates a high capacity for discrimination (Semwal et al., 2017). A three-stage process is used to analyze the extracted features. At first, the F-value is calculated for each of the 135 features using the ANOVA test. Next, the F-values of the features in each feature category are summed up. The last step in choosing the discriminative features is to sort the feature group F-values. The analysis of variance revealed that InJAF features are the least informative across all datasets. The experimental outcomes reflect the same pattern. Figure 5.12 compares the total F-values of feature groups for each gait event across datasets.

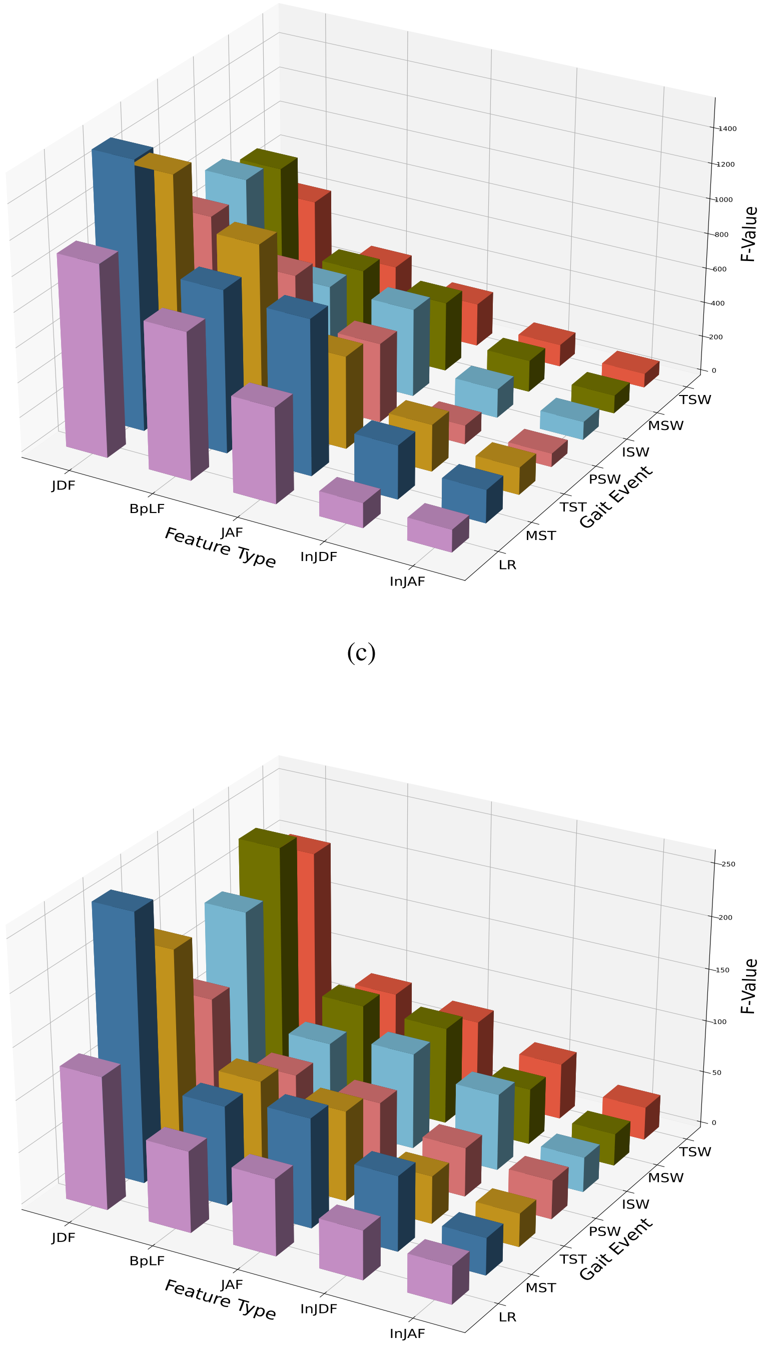


(a)



(b)

Figure 5.12: Comparison of total F-Value of feature groups in gait events. (a) UPCV1. (b) KGBD (continued in the next page).



(c)

(d)

Figure 5.12: Comparison of total F-Value of feature groups in gait events (continued from the previous page). (c) UPCV2. (d) KS20 (130°).

Analysis of Statistical Significance of the Proposed System

One-way analysis of variance test is performed on the results obtained by different experimental settings based on $Five - fold - cross - validation$ from K-folds on KGBD, UPCV1, and UPCV2 as in (Semwal et al., 2017) to determine the statistical significance of all the different experimental settings with different classifiers. One-way analysis of variance yields $P-value$ of 1.227e-25 for UPCV1, 9.2523e-27 for UPCV2, and 4.967e-53 for KGBD. The KS20 dataset was also subjected to this analysis, with the results obtained using $Cross - view - split$, and resulted in $P - value$ of 0.0034. The null hypothesis H_0 is considered as "average recognition Accuracy across all classifiers are equal". The results show that $P - Value < alpha$, with $alpha = 0.05$, for all datasets. As a result, the null hypothesis H_0 cannot be true.

5.1.2.9 Computational Cost Analysis and Discussion

Three components—feature extraction, score fusion, and an LSTM-based DL model contribute to the overall computational cost of the proposed system. The proposed system extracted gait event-specific features from a single frame and two frames that were consecutive. The complexity of feature extraction in terms of time is estimated to be $O(n)$, where n is the number of frames in the walking sequence. Estimated time complexity of score fusion is $O(L)$, where L represents the number of labels in the dataset. According to (Justus et al., 2018), Equation (5.19) gives the total execution time per epoch for a DL model.

$$E = pT_b, \tag{5.19}$$

Where, p is the total number of batches and T_b is the total estimated time to perform forward and backward passes on a single batch using the formula in Equation (5.20).

$$T_b = \sum_{i=0}^{l} b_{M(i)}, \tag{5.20}$$

Where, l, $b_{M(i)}$, and $M(i)$ denote the number of layers, i^{th} layer estimated batch execution time, and type of layer i, respectively.

The proposed DL model is comprised of two LSTM, BN, one Dense, and Softmax layers. The LSTM is spatially, and temporally local (Hochreiter and Schmidhuber, 1997). Consequently, input length does not affect network storage requirements (Tsironi et al., 2017). Hence, the estimated time complexity per timestamp per weight is $O(1)$. Thus the total time complexity per timestamp is $O(W_L)$, where W_L denote total weights that are approximately equal to total parameters in the LSTM layer. The BN primarily entails determining the mean and variance of each minibatch. Each operation's approximate time complexity per batch is $O(b)$, where b denotes input length. Similarly, the number of computations in a Dense layer is also proportional to its weights, and it is approximate $O(W_d)$, where W_d denotes the count of weights. Thus employing Equations (5.19) and (5.20), the total time required for forward and backward passes on a single epoch is estimated as defined in Equation (5.21).

$$Est = p * (2 * O(W_L) + 2 * O(b)) + 2 * O(W_d)), \tag{5.21}$$

The $Leave - one - sequence - out$ evaluation protocol employs all three subsystems. So the estimated time for e epochs is defined in Equation (5.22).

$$Est_{pt1} = O(n) + (p * (2 * O(W_L) + 2 * O(b)) + 2 * O(W_d))) * e + O(L), \tag{5.22}$$

The $Five - fold - cross - validation$, $Random - split$, and $Cross - view - split$ evaluation protocols does not involve score fusion. So, the total estimated time for e epochs is defined in Equation (5.23).

$$Est_{pt1} = O(n) + (p * (2 * O(W_L) + 2 * O(b)) + 2 * O(W_d))) * e, \tag{5.23}$$

The total number of computations needed is highly sensitive to both the feature count and the DL model parameters in these evaluation methods. The proposed LSTM-based DL model uses a significantly smaller number of parameters than other state-of-the-art methods. The DL model proposed in (Bari and Gavrilova, 2019) has 4494592 parameters, whereas the proposed system has only 189210 parameters, thereby decreasing the computational cost by approximately 24 times.

5.2 Attention-guided Residual Deep Learning Model and Gait Event-specific Features for Human Identification from Skeleton-based Gait Data

This study employs gait-event-specific features to analyze the temporal behavior of human limbs across various gait events to recognize human gait. Residual connections in the DL model are used in a wide variety of applications at the present time. In addition, the DL model uses several Attention models to focus on the most relevant aspects of a classification problem. Thus, this work proposes an improved DL model for HI, which uses residual connections and an Attention module. The key contributions of this thesis work are listed below.

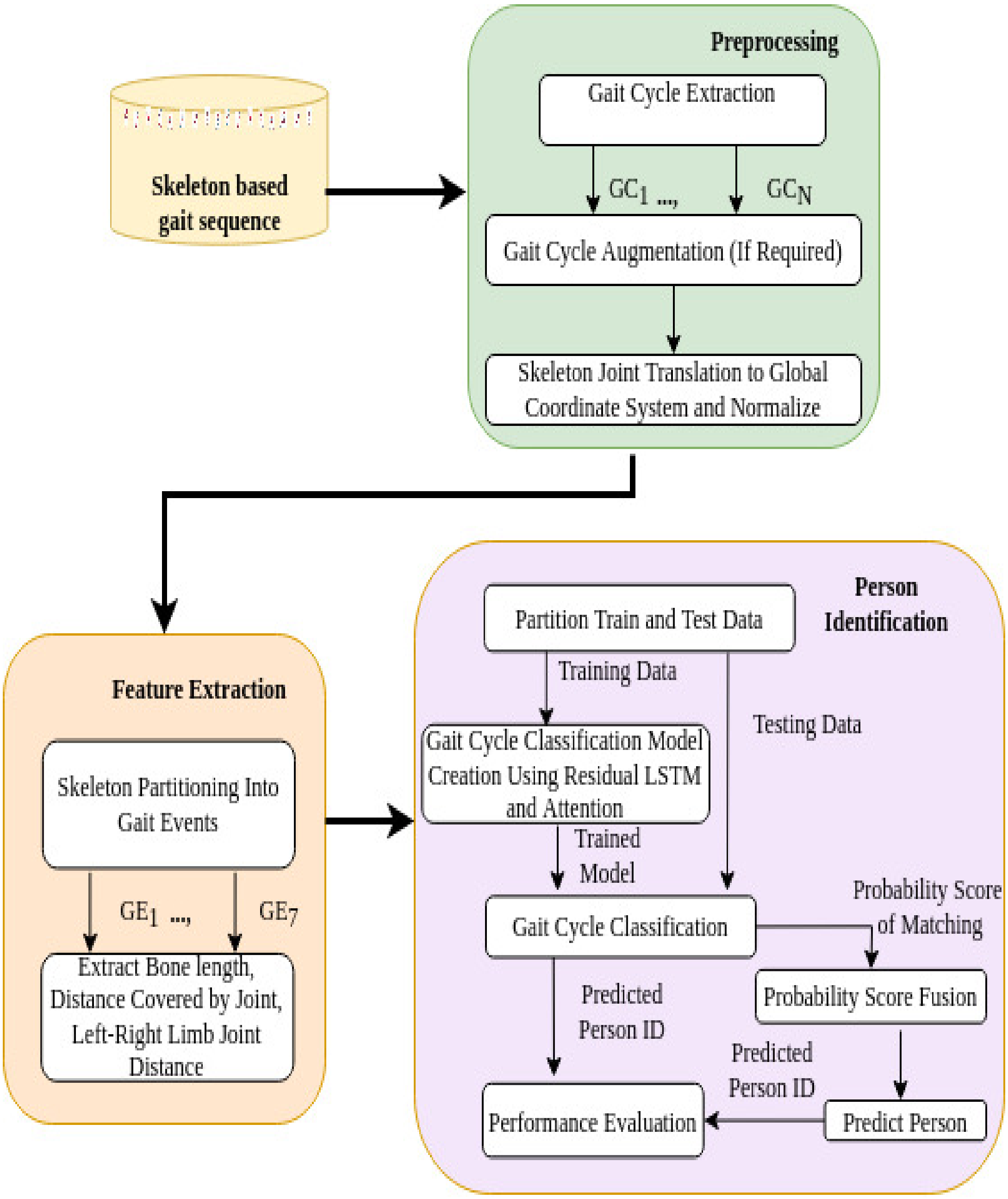
Contributions:

- Concentrating feature extraction efforts on the individual gait event streams constituting the gait cycle.
- Proposed novel, robust features of gait events contributing to the whole gait cycle. We also use a quantitative summary of the features to reduce the total number of features drastically.
- Proposed novel Attention-guided residual LSTM/GRU-based DL models that effectively capture the most discriminant gait event features for gait recognition.
- Primary emphasis on reducing the computation complexity of the overall approach by reducing the number of features and complexity of the DL model without sacrificing the system performance.
- Evaluated the performance of the proposed method using benchmark evaluation metrics and protocols on single and multi-view skeleton-based gait datasets and compared it to state-of-the-art approaches.

5.2.1 Proposed Methodology

As part of the proposed work, the gait cycles of a walking sequence are detected and preprocessed. Then, the features of each gait event are extracted. Moreover, an Attention-guided residual LSTM/GRU-based DL model is proposed to learn the extracted features and identify the person. In addition, the human is identified by combining the results of gait cycle recognition and a score fusion operation. The various phases of the proposed work are depicted in Figure 5.13. Pre-processing, Feature Extraction, and Person

Identification are the primary subsystems of the proposed system. Following are the specifics of each subsystem.



Skeleton based gait sequence
Preprocessing
Gait Cycle Extraction
GC_1 ...,
GC_N
Gait Cycle Augmentation (If Required)
Skeleton Joint Translation to Global Coordinate System and Normalize
Feature Extraction
Skeleton Partitioning Into Gait Events
GE_1 ...,
GE_7
Extract Bone length, Distance Covered by Joint, Left-Right Limb Joint Distance
Person Identification
Partition Train and Test Data
Training Data
Testing Data
Gait Cycle Classification Model Creation Using Residual LSTM and Attention
Trained Model
Gait Cycle Classification
Probability Score of Matching
Probability Score Fusion
Predicted Person ID
Predicted Person ID
Performance Evaluation
Predict Person

Figure 5.13: Proposed human identification system using gait event-specific features and residual DL model with Attention.

5.2.1.1 Pre-processing

This involves primarily gait cycle extraction, an optional data augmentation. In addition, all the skeleton joints are transformed into a global coordinate system and normalized. The details of all these sub-tasks are already explained in subsection 5.1.1.1.

5.2.1.2 Feature Extraction

This work proposes three gait-event-specific features, namely: Mean Length of Bones (MLB), Distance covered by Joint (DistJ), and Mean of Distance between Joints in Left and Right body part (MJDLR). The features considered are shown in Figure 5.14.

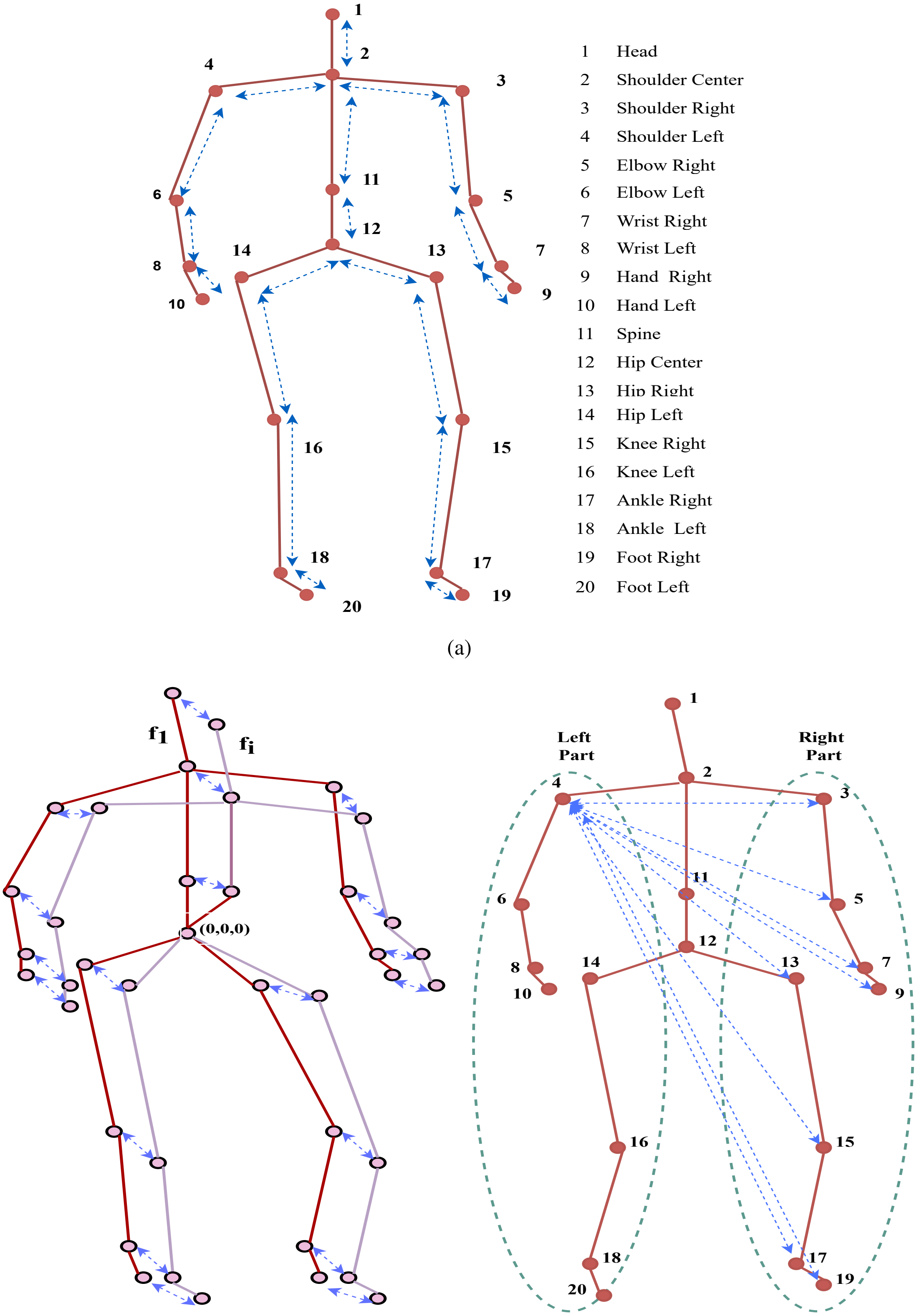
For MLB feature extraction, the length of each bone in the skeleton frame belonging to a gait event is initially determined. As bone length must be static, the next step is to compute the mean of each bone's length over the gait event. MLB contributed 19 features to the feature vector (Figure 5.14 (a)). DistJ computes the total distance covered by each joint during the gait event and contributes 19 features to the feature vector (Figure 5.14 (b)). For MJDLR, the whole skeleton is vertically partitioned into three parts. As the distance between joints in the body's left and right parts varies, we initially computed this distance. Next, we determined the average distance between each joint on the left side of the body and each joint on the right side. Figure 5.14 (c) depicts the distance between the left shoulder and all right-side joints. As a result, 102×7 features were extracted for each gait cycle.

5.2.1.3 Person Identification

This subsystem begins by partitioning the feature set according to the evaluation methods. The proposed DL model is used to train the features for gait cycle classification. Score fusion is performed based on the requirements of evaluation methods.

Proposed Residual DL Model with Attention

This section discusses the Attention-guided residual LSTM/GRU based DL model proposed to learn the temporal relationship between extracted features. Attention mechanism is used in the proposed model so that it focuses primarily on the essential gait features. The proposed DL model is illustrated in Figure 5.15.



(a)

(b) (c)

Figure 5.14: Features extraction. (a) MLB. (b) DistJ. (c) MJDLR.

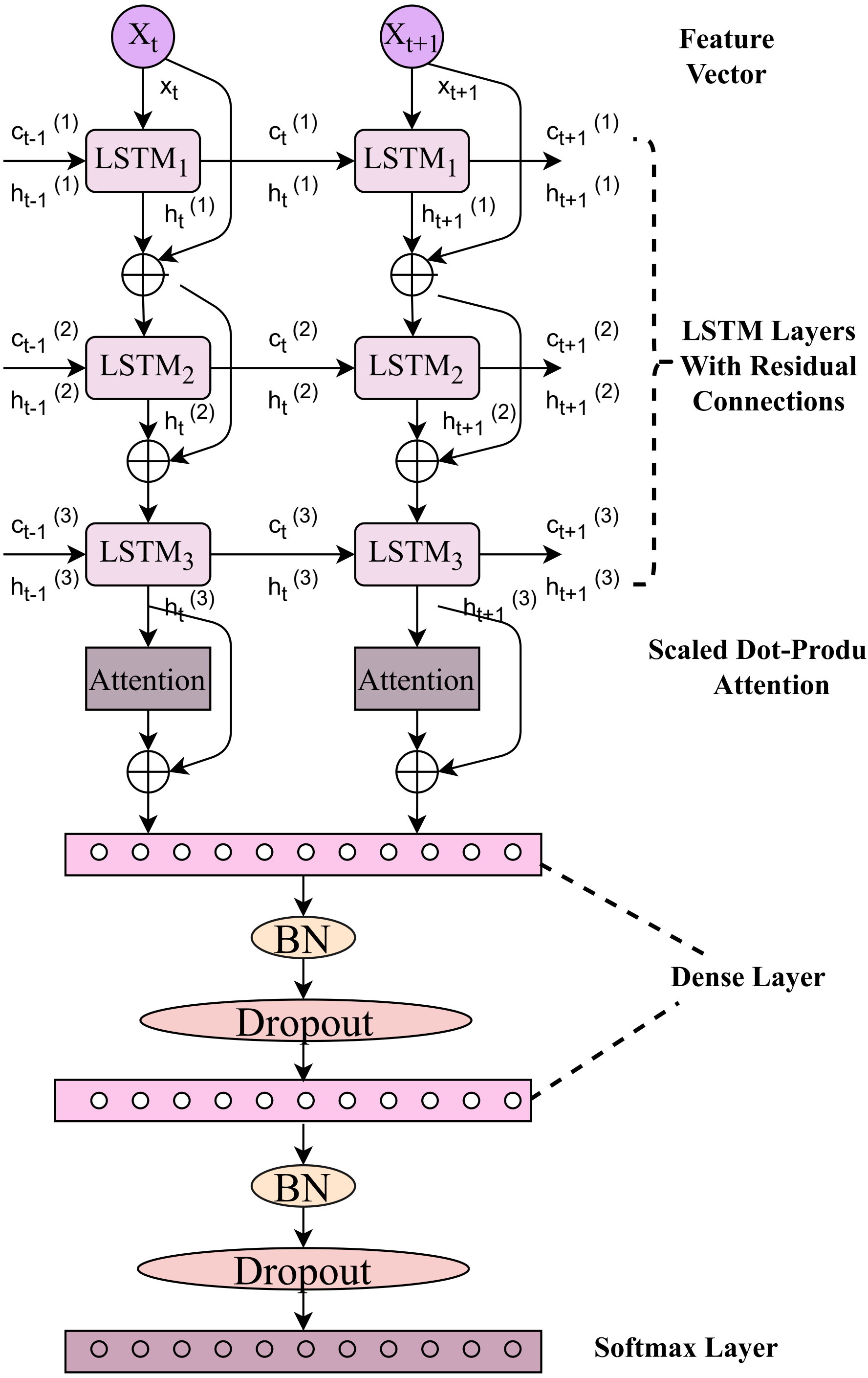


Figure 5.15: Proposed Attention guided residual LSTM.

Attention Mechanism

Attention is added after the final LSTM/GRU layer to focus on more important features. In the proposed method, a variant of the Self-Attention algorithm is used to generate the context-specific feature representation by correlating the temporal features extracted by

LSTM/GRU. Typically, the Attention unit receives three input vectors: Queries (Q), Keys (K), and Values (V). The Attention unit maps the Q and a set of Key-Value (K-V) pairs to an output sequence. Attention computes the weighted sum of the Values as the output, where the weight assigned to each V is determined by an alignment function between Q and K. In the proposed approach, the "Scaled Dot-Product Attention" (Vaswani et al., 2017) function is used to direct the residual LSTM/GRU model to prioritize context-specific temporal gait features. At first, the "Scaled-dot-product" Attention computes the dot products of Q and K and divides it by the scaling factor $\sqrt{dk}$ to prevent an excessively large result. Here, dk represents the dimension of the query and key vectors. A Softmax function is then applied to normalize the result and the normalized result is multiplied by V to gather weights on values. The overall computation procedure is as shown in Equations (5.24) and (5.25). This section discusses the proposed Attention-guided residual LSTM/GRU DL model to learn the temporal relationship between extracted features. Attention mechanism is used in the proposed model so that it focuses primarily on the essential gait features.

$$Attention(Q, K, V) = Softmax(\alpha)V \tag{5.24}$$

$$\alpha = \frac{QK^T}{\sqrt{dk}} \tag{5.25}$$

Further, two Dense layers with $tanh$ activation process the concatenated context vector from the Attention unit and the output h_t of the final LSTM/GRU layer. BN layer normalizes Dense layer output. Dropouts are added after the BN layer to prevent network over-fitting. A Softmax layer calculates the probability score for matching processed features to person Ids.

Probability Score Fusion

An optional probability score fusion function combines the probability matching score of gait cycles from the same walking sequence to identify the person based on all their scores. The proposed method followed score fusion operation discussed in Section 5.1.1.7 (Page No. 86) to combine the gait cycle recognition scores from entire walking sequence to identify the person.

5.2.2 Experiments, Results, and Analysis

Several experiments were performed using the residual 3-Layer, 2-Layer, and 1-Layer LSTM and GRU models with and without Attention units. In addition, we tested a variety of machine learning-based classification models including KNN, SVM, and RF, using the same set of features.

5.2.2.1 Dataset

The proposed approach is evaluated based on five 3D skeleton-based gait datasets, UPCV1, UPCV2, KGBD, IAS-Lab RGBD-ID (IAS-Lab) (Munaro et al., 2014; Nanni et al., 2016), and KS20 dataset. Four of these datasets are the single view, and the fifth dataset is multi-view. The details of UPCV1, UPCV2, KGBD, KS20 are already given in Section 5.1.2.1 (Page No. 88). The additional dataset used in this work is IAS-Lab.

IAS-Lab (Munaro et al., 2014) is a multi-modal gait dataset consisting of 11 individuals. In this work, only the 3D skeleton data provided by the dataset was utilized. 'Training', 'TestingA', and 'TestingB' are the three subsets of data provided by the dataset. Sequences of 'TestingA' were recorded with participants wearing distinct clothing than 'Training'. Whereas, 'TestingB' was collected in a separate room while wearing the same clothing as 'Training'. The statistics of all these datasets are provided in Table 5.13.

Table 5.13: Statistics of skeleton-based gait dataset used in the proposed work.

	UPCV1	**UPCV2**	**KGBD**	**KS20**	**IAS-Lab**
No. of Subjects	30	30	164	20	11
No. of Video Sequence	150	300	822	300	11+11+11
Multi-View	No	No	No	Yes	No

5.2.2.2 Evaluation Protocols and Metrics

The evaluation protocols mentioned in Section 5.1.2.2 (Page No. 89) are used to evaluate the proposed work. They are namely: $Leave - one - sequence - out$, $Five - fold - cross - validation$ on UPCV1, UPCV2, and KGBD, and $Cross - view - split$, $Random - split$ on KS20. For IAS-Lab, as the dataset provides separate training and testing sets, the same is used for evaluation.

The same set of evaluation metrics, namely: $Precision$, $Recall$, $Accuracy$, $F-Score$, and CMC test curves, are used to evaluate the proposed work. The details of these are already given in Section 5.1.2.3 (Page No. 90).

5.2.2.3 Training Parameters

The models are trained for 200-450 epochs based on dataset requirement. We employed a 64-batch size and a learning rate of 0.001 to train the models. In addition, the learning rate is reduced by a factor of 0.70 to improve model's performance. The model is trained with Categorical Cross-Entropy loss function, and Adam optimizer.

5.2.2.4 Experiments Considering Single-view Scenario

In this, we considered the entire dataset and divided the samples into training and testing based on evaluation protocols. The direction of data capture is not considered in the partition process.

Table 5.14: Performance of proposed residual deep learning model on KGBD, UPCV1, and UPCV2 based on $Five-fold-cross-validation$.

Approach	Methods	Dataset											
		KGBD				UPCV1				UPCV2			
		Precision	Recall	F-Score	Accuracy	Precision	Recall	F-Score	Accuracy	Precision	Recall	F-Score	Accuracy
DL Models	1 layer LSTM + Attention	96.6	96.2	96.2	96.28	93.8	95.2	94	95.28	**100**	99.4	99.4	99.67
	2 layer LSTM + Attention	96.2	96	96	96.092	93.6	94.6	93.4	94.70	**100**	99.4	99.4	99.67
	3 layer LSTM + Attention	**97**	**96.8**	**96.8**	**96.88**	**95.6**	**97**	**96.2**	**97.06**	**100**	**99.6**	**99.6**	**99.78**
	1 layer GRU + Attention	96.2	95.6	95.6	95.82	92.6	94.6	93.2	94.7	99.8	99.4	99.4	99.56
	2 layer GRU + Attention	95.4	95.4	95.4	95.37	92.8	94.6	93.2	94.69	99.8	99.4	99.4	99.56
	3 layer GRU + Attention	95.8	95.2	95.2	95.49	94	95.8	94.6	95.87	**100**	99.2	99.2	99.56
ML Approaches	SVM	91.6	91.2	91.2	91.2	77.8	82.6	79	82.6	97.6	97.4	97.4	97.4
	KNN-1	80.2	79.2	78.4	79.2	83.4	86.8	84	86.8	94.4	93.8	93.6	93.8
	KNN-5	76.4	72.8	72	72.8	56	61.4	55.4	61.4	90.6	90	89.6	90
	RF-10	70.2	66.4	66.4	66.4	73.2	78.6	74	78.6	96.2	95.6	95.6	95.6
	RF-5	60	49.2	50.2	49.2	59.6	64.2	59.4	64.2	92	90.8	90.6	90.8

The results obtained for KGBD, UPCV1, and UPCV2 using both DL models and ML classifications based on $Five - fold - cross - validation$ are shown in Table 5.14. On both KGBD and UPCV1, the proposed model with 3-Layer residual LSTM and Attention performs better than other models. On UPCV2, the performance of all DL models was nearly identical, with minor variations. On each of the three datasets, classifications based on machine learning performed poorly. In KNN and RF, we repeated the experiment with different K values and estimator counts.

Table 5.15: Performance of proposed residual deep learning model on KGBD and UPCV2 using $Leave - one - sequence - out$ protocol.

Method	**Dataset**							
	KGBD				**UPCV2**			
	Precision	**Recall**	**F-Score**	**Accuracy**	**Precision**	**Recall**	**F-Score**	**Accuracy**
1 layer LSTM + Attention	100	98	98.6	97.92	100	99.4	99.6	99.33
2 layer LSTM + Attention	100	98.4	99	**98.4**	100	99.7	99.8	**99.66**
3 layer LSTM + Attention	100	98	98.8	98.04	100	99.4	99.6	99.33
1 layer GRU + Attention	100	98.2	98.6	98.29	100	99.7	99.8	**99.66**
2 layer GRU + Attention	100	98.4	99	**98.4**	100	99.7	99.8	**99.66**
3 layer GRU + Attention	100	97.8	98.6	97.8	100	99.4	99.6	99.33

Table 5.15 reports the results obtained for $Leave - one - sequence - out$ protocol on KGBD and UPCV2. As we use probability scores to perform score fusion, we did not conduct ML algorithm-based experiments. The outcomes demonstrate that 2-Layer GRU and LSTM models with Attention outperformed other models.

Table 5.16 displays the results obtained on KS20 by employing the $Random - split$ protocol. It is observed that the proposed 3-Layer residual LSTM with Attention achieved a 91% $Accuracy$. With an $Accuracy$ of 92%, the 3-Layer GRU model is superior. Other models, meanwhile, performed poorly. In addition, the ML models did not produce the expected outcomes. Table 5.16 also reports the results on IAS-Lab ‘TestingA’ and ‘TestingB’. With a score of 58.33%, 2-Layer GRU with Attention demonstrated the highest $Accuracy$ on 'TestingA'. Comparatively, the $Accuracy$ of the 3-Layer LSTM with the Attention model was 56.25%. On 'TestingB', 3-Layer residual LSTM with Attention secured the highest $Accuracy$ of 60%. However, machine-learning-based models demonstrated poor performance.

Table 5.16: Performance of proposed residual deep learning model on KS20 using $Random-split$ protocol, and on IAS-Lab the testing results on 'TestingA' and 'TestingB'

Approach	**Methods**	**Dataset**											
		KS20				**IAS-TestA**				**IAS-TestB**			
		Precision	**Recall**	**F-Score**	**Accuracy**	**Precision**	**Recall**	**F-Score**	**Accuracy**	**Precision**	**Recall**	**F-Score**	**Accuracy**
DL models	1 layer LSTM + Attention	83	78	78	77.89	59	50	48	50	43	49	44	48.88
	2 layer LSTM + Attention	85	81	81	81	54	44	43	43.75	50	56	49	55.55
	3 layer LSTM + Attention	**93**	91	91	91	64	56	54	56.25	**53**	**60**	**54**	**60**
	1 layer GRU + Attention	80	79	79	78.95	55	54	50	54.16	48	53	49	53.33
	2 layer GRU + Attention	86	83	83	83	**74**	**58**	**59**	**58.33**	42	47	43	46.6
	3 layer GRU + Attention	**93**	**92**	**92**	**92**	49	52	49	52.08	54	53	47	53.33
ML approaches	SVM	74	71	70	71	30	31	30	31	46	42	38	42
	KNN-1	61	56	55	56	36	25	26	25	37	36	32	36
	KNN-5	52	50	47	50	30	21	21	21	31	29	25	29
	RF-10	47	41	39	41	29	29	27	29	49	49	45	49
	RF-5	59	58	57	58	29	27	26	27	49	49	45	49

Comparison of proposed work with existing methods

Table 5.17 shows the comparison of Rank-1 $Accuracy$ of proposed models with state-of-the-art works employing the same evaluation protocols. The first four rows' results are based on those published in (Bari and Gavrilova, 2019). The proposed 3-Layer residual LSTM with Attention outperformed all other works on the UPCV1 dataset. On the other hand, the proposed work on KGBD demonstrated 1.18%, 2.43% less Rank-1 $Accuracy$ than (Bari and Gavrilova, 2019), and (Bari and Gavrilova, 2022) respectively. However, the method (Bari and Gavrilova, 2019) is computationally intensive with a DL model with extremely high neural network parameters, i.e., 4494592 total parameters, without considering the Softmax layer. Also, (Bari and Gavrilova, 2022) has 471902, 480612 total parameters on UPCV1 and KGBD, respectively. In contrast, the proposed deep learning model with 3-Layer residual LSTM and Attention (the proposed model with the highest number of parameters) has only 335172 total parameters without considering the Softmax layer. As the proposed 3-layer residual LSTM model with Attention has 335172 total parameters when compared to that of (Bari and Gavrilova, 2019) which has 4494592 total parameters, hence the overall computational cost is reduced by approximately 14 times.

Table 5.17: Comparison of Rank-1 $Accuracy$ (%) with state-of-the-art-works on $Five-fold-cross-validation$.

Method	Dataset	
	KGBD	**UPCV1**
Unsupervised Clustering (Ball et al., 2012)[a]	37.55	57.0
Gait-Skeleton. (Preis et al., 2012)[a]	75.46	78.0
Viewinvariant-Skeleton. (Sun et al., 2018)[a]	79.76	82.67
Relative distance (Yang et al., 2016)[a]	94.88	86.67
ANN (Bari and Gavrilova, 2019)	98.08	95.30
KinectGaitNet (Bari and Gavrilova, 2022)	**99.33**	96.91
3 Layer Residual LSTM + Attention (Proposed)	96.9	**97.06**

[a] Are based on published results in (Bari and Gavrilova, 2019).

Table 5.18 provides the detailed comparison results of proposed work with state-of-the-art works using the same evaluation protocol. On the IAS-Lab, the results of ‘TestingA’ and ‘TestingB’ are compared to those of a number of other studies. The proposed work achieved 1.7%, and 2.5% lower Rank-1 $Accuracy$ on ‘TestingA’ and ‘TestingB’, respectively, compared to (Rao et al., 2021). In contrast to (Rao et al., 2021), the proposed work achieved superior performance on KS20 using the $Random-split$ protocol. Using the $Leave-one-sequence-out$ protocol, the proposed work achieved 7.8% greater Rank-1 $Accuracy$ than (Rao et al., 2021) on KGBD. The performance is 0.4% better than the best performance recorded in (Li et al., 2017) on KGBD. In addition, on UPCV2 using the $Leave-one-sequence-out$ evaluation protocol, the proposed approach demonstrated superior performance in comparison with existing state-of-the-art works. Overall, the proposed work outperformed several state-of-the-art works on various datasets based on various evaluation protocols.

5.2.2.5 Experiments Considering Multi-view Scenario

Table 5.19 gives the details about the results obtained by the proposed work on the KS20 dataset in the $Cross-view-split$ evaluation protocol. In addition, Table 5.19

Table 5.18: Comparison of Rank-1 $Accuracy$ (%) of proposed work with state-of-the-art works: on IAS-Lab: two testing sets 'TestingA' and 'TestingB', on KS20: $Random - split$, on KGBD and UPCV2: $Leave - one - sequence - out$

Method Id	**Method**	**IAS-TestA**	**IAS-TestB**	**KS20**	**KGBD**	**UPCV2**
S_1	Gait_Anthro (Andersson and Araujo, 2015)	-	-	-	87.7	-
S_2	Skeleton +single LSTM (Haque et al., 2016)	20.0	19.1	-	-	-
S_3	Information Fustion (Kastaniotis et al., 2016)	-	-	-	-	97.05
S_4	Context_aware (Nambiar et al., 2017a)	-	-	88.67	-	-
S_5	Context_unaware (Nambiar et al., 2017a)	-	-	79.33	-	
S_6	Posture based gait (Khamsemanan et al., 2018)	-	-	-	97.5	-
S_7	Dynamic LSTM (Li et al., 2017)	-	-	-	96.56	-
S_8	SKeGEI+DA+CNN-LSTM (Liu et al., 2019)	-	-	-	97.39	
S_9	joint distance+angle+CNN (Huynh-The et al., 2020)	-	-	87.63		99.65
S_10	PoseGait (Liao et al., 2020)	41.4	**37.1**	70.5	90.6	-
S_11	gait encoding (Rao et al., 2020)	56.1	58.2	-	87.7	-
S_12	Geometric ConvNet (Hosni and Amor, 2020)	-	-	-	-	92.41
S_13	Self Supervised (Rev. Rec) (Rao et al., 2021)	**60.1**	**62.5**	86.9	86.9	-
S_14	Self Supervised (Rev. Rec.Plus) (Rao et al., 2021)	59.1	62.2	**92.0**	90.6	-
S_15	Adversarial Adversarial (Chen et al., 2022)	61.1	63.9	88.0	87.4	-
S_16	3 Layer LSTM + Attention (Proposed)	56.25	60	91	98.04	99.33
S_17	2 Layer LSTM + Attention (Proposed)	43.75	55.55	81	**98.4**	**99.66**
S_18	3 Layer GRU + Attention (Proposed)	52.08	53.33	**92**	97.8	99.33
S_19	2 Layer GRU + Attention (Proposed)	58.33	46.6	83	**98.4**	**99.66**

Method_Id S_16 to S_19 are the part of proposed work

compares the Rank-1 $Accuracy$ with several other existing works. In this (Liao et al., 2020) results are based on (Rao et al., 2021). The proposed work achieved superior performance on three different views than the existing works. Also, the overall average among all the views achieved by the proposed work is greater than all other published results of the existing state-of-the-art works.

5.2.2.6 Ablation Experiments

As part of the ablation study, numerous experiments were conducted. The importance of the Attention module is demonstrated by conducting several experiments without

Table 5.19: Rank-1 $Accuracy$ (%) on KS20 using $Cross - view - spilt$ and comparison with state-of-the-art works.

Method	0°	30°	90°	130°	180°	**Average**
Pose Gait, (Liao et al., 2020)	24.6	19.1	29.7	27.3	25	25.14
Self Supervised (Rev.Rec) (Rao et al., 2021)	44.4	54.9	55.0	41.9	53.4	49.92
Self Supervised (Rev.Rec.Plus) (Rao et al., 2021)	**48.8**	53.6	54.9	44.5	**57.5**	51.86
1-Layer LSTM + Attention (Proposed)	35.59	50.84	83.05	69.49	50.84	57.962
2-Layer LSTM + Attention (Proposed)	33.89	59.32	81.35	72.88	40.67	57.622
3-Layer LSTM + Attention (Proposed)	35.59	**62.71**	**84.75**	71.18	55.93	**62.032**
1-Layer GRU + Attention (Proposed)	37.28	55.9	**84.75**	74.57	42.37	58.96
2-Layer GRU + Attention (Proposed)	37.28	52.45	83.05	**76.27**	44.06	58.62
3-Layer GRU + Attention (Proposed)	37.28	62.71	83.05	**76.27**	42.37	60.33

Table 5.20: Rank-1 $Accuracy$ (%) of the different residual DL models without Attention module using $Five - fold - cross - validation$.

Method	**Dataset**					
	KGBD	**IAS-TestA**	**IAS_TestB**	**KS20**	**UPCV1**	**UPCV2**
2-Layer LSTM	96.2	47.91	51.1	80	93.52	99.4
3-Layer LSTM	96.03	54.17	40	87	93.52	99.6
2-Layer GRU	95.7	52.08	51.1	88	94.09	99.3
3-Layer GRU	95.5	50	44.4	81	92.35	99.0

Table 5.21: Rank-1 $Accuracy$ (%) of the different residual deep learning model on KS20 using $Cross - view - split$ protocol without Attention module.

Method	0°	30°	90°	130°	180°	**Average**
2-Layer LSTM	37.28	54.24	77.97	67.8	38.98	55.25
3-Layer LSTM	37.28	49.15	77.97	69.49	45.76	55.93
2-Layer GRU	38.98	57.63	86.44	72.88	45.76	60.33
3-Layer GRU	35.59	57.63	84.7	69.49	37.29	56.94

the Attention module. Table 5.20 details the gait recognition performance of various models lacking an Attention unit on the KGBD, UPCV1, UPCV2, KS20, and IAS-Lab

Table 5.22: Rank-1 $Accuracy$ (%) of proposed residual DL model on KS20 without data augmentation.

Method	**Random-split**	**Cross-view-split**					
		0^0	30^0	90^0	130^0	180^0	**Average**
2-Layer LSTM + Attention	63	32.2	54.23	83.05	76.27	44	57.95
3-Layer LSTM + Attention	70	28.81	59.32	83.05	62.71	40.67	54.91
2-Layer GRU + Attention	62.1	33.89	49.15	86.44	62.71	38.98	54.23
3-Layer GRU + Attention	66	37.28	61.017	74.57	64.4	38.98	55.25

datasets. In addition, Table 5.21 displays the outcomes of the KS20 $Cross - view - split$. Compared to the results of the proposed work discussed earlier with Attention, the models without Attention performed significantly worse.

In addition, several experiments were conducted on KS20 without the use of augmented data. As the walking sequence of the KS20 dataset only contains a single gait cycle, the proposed data augmentation increases the training data size of the KS20 dataset. Table 5.22 shows the Rank-1 results obtained using the $Random - split$ and $Cross - view - split$ evaluation protocols. It is revealed from the experiments that those without augmentation got a lower average Rank-1 $Accuracy$ than their counterparts with augmentation shown in Table 5.19 for the $Cross - view - split$ evaluation protocol. However, the model with 2-Layer LSTM with Attention without augmented data showed a small higher $Accuracy$ by 0.3% than its counterpart in Table 5.19. As shown in Table 5.22, the experiments without augmentation for the $Random - split$ protocol yielded inferior results compared to their counterparts with augmentation. The proposed data augmentation has proven its efficacy in gait recognition.

The ablation experiments were conducted on the feature set as well. To illustrate the importance of each feature type in gait recognition in the proposed approach, we conducted a series of experiments using both individual features and various combinations of features. Table 5.23 displays the Rank-1 gait recognition $Accuracy$ for the IAS-Lab, KS20 with $Random - split$, and KGBD, UPCV1, and UPCV2 with 5-Fold cross-validation with the various feature set combination. The Rank-1 $Accuracy$ for the KS20 $Cross - view - split$ protocol with various feature sets is displayed in Table

5.24. It is observed from Tables 5.23 and 5.24 that optimal performance is achieved by combining all three types of features. However, we did find an outlier on ‘TestingB’ of IAS-Lab. In this 3-Layer Attention-based residual LSTM, the MJDLR feature type achieved 62.22% Rank-1 $Accuracy$, while combining three features yielded 60% $Accuracy$. Furthermore, the MJDLR obtained 54.16% on ‘TestingA’ and 57.77% on ‘TestingB’ when using 3-Layer residual GRU with Attention. In contrast, when all feature types are used together, the results are only modest (52.08% on ‘TestingA’ and 53.33% on ‘TestingB’). Similarly, we found that the average $Accuracy$ across all views is better with the combination of all the feature sets, despite slightly poor performance on some views on KS20 using the $Cross-view-split$ protocol using all the features.

Table 5.23: Rank-1 $Accuracy$ (%) using different feature combinations using $Five-fold-cross-validation$

Feature Set	**Method**	**Dataset**					
		KGBD	**IAS-TestA**	**IAS-TestB**	**KS20**	**UPCV1**	**UPCV2**
MLB	3-Layer LSTM + Attention	62.7	31.25	33.33	73	63.52	98.26
DistJ		32.35	35.41	17.77	20	59.40	83.99
MJDLR		91.85	56.25	62.22	74.74	92.34	99.43
MLB+DistJ		84.47	35.41	26.66	73	87.64	99.45
MLB+MJDLR		96.34	56.25	17.77	80	94.70	99.56
DistJ+MJDLR		93.41	52.08	53.33	75.79	93.52	99.45
MLB+DistJ+MJDLR		96.88	56.25	60	91	97.06	99.78
MLB	3-Layer GRU +Attention	60.22	29.17	31.11	73	67.05	97.72
DistJ		33.01	31.25	15.55	76.84	55.29	82.7
MJDLR		90.40	54.16	57.77	22	90.58	99.23
MLB+DistJ		84.13	33.33	20	61.05	85.87	99.34
MLB+MJDLR		95.63	41.66	42	77	94.11	99.34
DistJ+MJDLR		93.04	47.91	46.66	78.95	94.69	99.45
MLB+DistJ+MJDLR		95.8	52.08	53.33	92	95.87	99.56

5.2.2.7 Cumulative Match Characteristic Curve

The CMC curves obtained on gait recognition are shown in Figure 5.16. The Figures 5.16 (a), (b), and (c) illustrate the CMC curves obtained on KGBD, UPCV1, and UPCV2, respectively using $Five-fold-cross-validation$. It is observed that

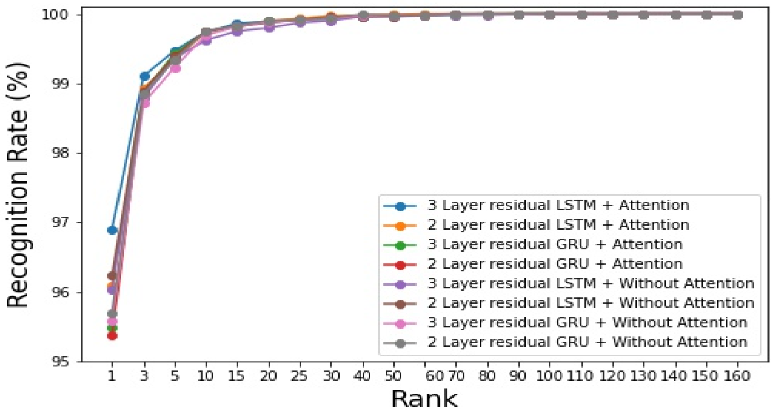
Table 5.24: Rank-1 $Accuracy$ (%) on different feature combinations on KS20 using $Cross-view-Split$

Feature Set	Method	0°	30°	90°	130°	180°	**Average**
MLB	3-Layer LSTM + Attention	25.42	50.4	64.4	64.4	37.8	48.38
DistJ		20.33	13.55	18.64	20.33	16.94	17.95
MJDLR		30.5	54.23	71.18	83.05	33.89	54.57
MLB+DistJ		16.94	45.76	64.4	54.23	27.11	41.68
MLB+MJDLR		30.5	59.32	81.35	83.05	37.08	58.3
DistJ+MJDLR		28.81	61.01	79.66	71.18	38.98	55.92
MLB+DistJ+MJDLR		35.59	62.71	84.75	71.18	55.33	62.03
MLB	3-Layer GRU +Attention	16.95	47.46	66.1	62.71	38.98	46.44
DistJ		15.25	16.95	16.95	22.03	22.03	18.64
MJDLR		35.59	52.54	79.66	79.66	42.37	57.96
MLB+DistJ		18.64	49.15	72.88	55.93	45.76	48.47
MLB+MJDLR		30.05	55.93	83.05	79.66	42.37	58.30
DistJ+MJDLR		35.59	61.01	86.44	74.57	37.28	58.97
MLB+DistJ+MJDLR		37.28	62.71	83.05	76.27	42.37	60.33

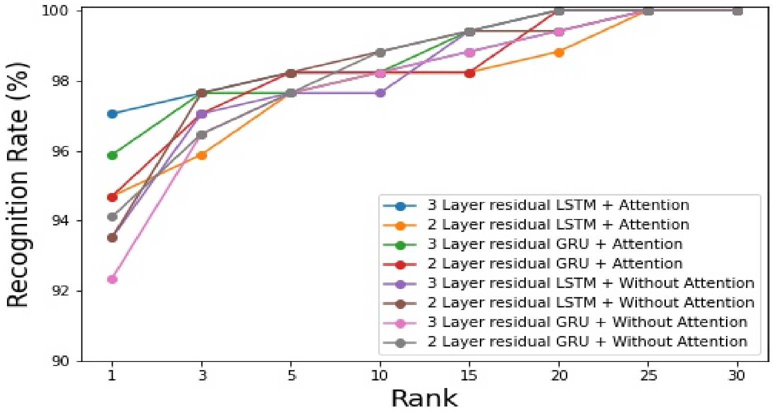
among all these, the proposed 3-Layer LSTM with Attention scored the highest Rank-1 $Accuracy$ than others. Figures 5.16 (d) and (e) are the CMC curves obtained for 'TestingA' and 'TestingB' of IAS-Lab dataset. Figure 5.16 (e) depicts the CMC curve for KS20 on $Random-split$. Here, the 2-layer LSTM Rank-1 $Accuracy$ is slightly lower than the GRU model. Overall, proposed 2 and 3-Layer GRU and LSTM models with Attention achieved higher accuracy (more than 99%) in lower ranks itself. Also, it is observed that the 3-Layer residual LSTM with Attention is superior in most of the experimental setups.

5.2.2.8 Further Statistical Analysis

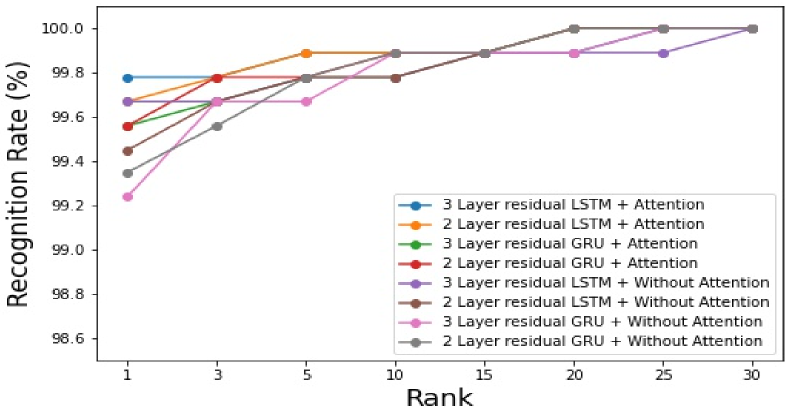
The statistical significance of the proposed DL model is evaluated by conducting ANOVA test on the recognition $Accuracy$ from set of folds as described in (Semwal et al., 2017). With significance level $\alpha = 0.05$ and null hypothesis H_0 is assumed as "The averages of Rank-1 Accuracy obtained on K folds in all experiments are equal", and the ANOVA test yielded P-values of 0, 0, and 1.22125e-15 for KGBD, UPCV2, and UPCV1, re-



(a)

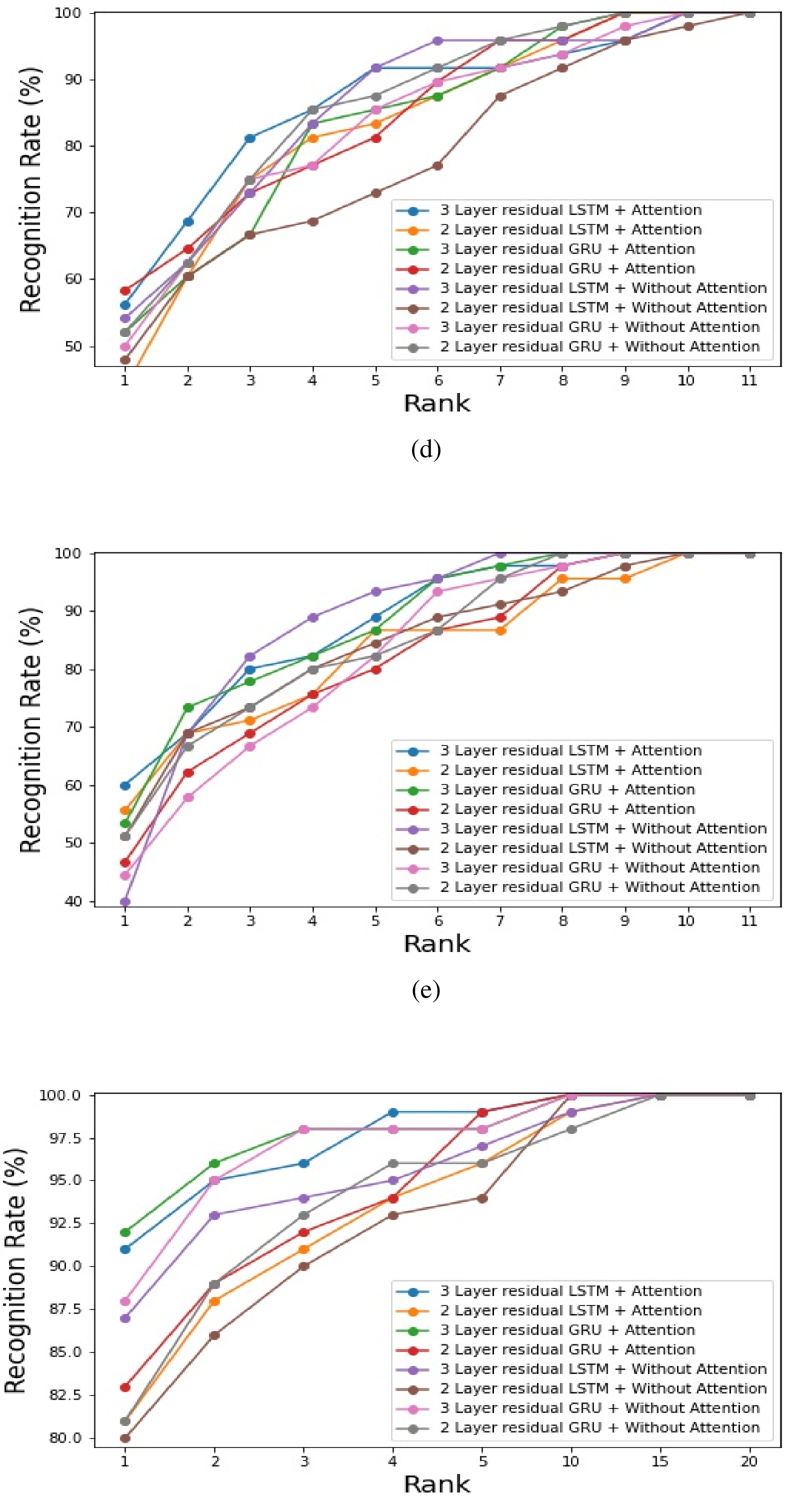


(b)



(c)

Figure 5.16: CMC curves obtained. (a) KGBD. (b) UPCV1. (c) UPCV2 (continued in the next page).



(d)

(e)

(f)

Figure 5.16: CMC curves obtained (continued from the previous page). (d) IAS-Lab TestingA. (e) IAS-Lab TestingB. (f) KS20.

spectively. The measured P-value is less than the significance threshold α. The null hypothesis H_0 is therefore rejected.

5.2.2.9 Computational Complexity Analysis

The proposed system extracts a quantitative summary of three types of gait event-specific features. The complexity of extracting MLB, and DistJ is $O(N \times V)$. To extract MJDLR, we initially extracted the distance between each skeleton joint in the left and right parts of the body. Hence, the computational complexity is $O(N \times (\frac{V}{2})^2)$. But, the number of skeleton joints considered is a constant (20). Thus, the overall computations required for feature extraction is $3 \times O(N)$. Where, N denotes the count of frames that make up one gait cycle. In this overall time complexity of the the proposed approach is estimated considering the proposed residual DL model comprising three LSTM layers with skip connection, two Dense layers with dropout and BatchNormalization, and an Attention unit followed by a Softmax layer. Referring to Equations (5.19) and (5.20), for each epoch, the computation complexity is estimated by using the Equation (5.26).

$$T_{fb} = N_b * (3 * O(W_L) + 2 * O(I_b) + O(W_a) + 3 * O(W_d)), \tag{5.26}$$

Thus, for E epochs, the time complexity of the proposed residual DL model based on LSTM and Attention is defined by the Equation (5.27). Where, W_L is the total count of weights from LSTM layers that roughly equals to count of parameters, I_b denotes the length of the batch input, W_d denotes the count of weights in Dense layer, and W_a is the total parameters in Attention.

$$\begin{aligned} Est_{pt1} = O(N) + E * (N_b * (3 * O(W_L) + \\ 2 * O(I_b) + O(W_a) + 3 * O(W_d))), \end{aligned} \tag{5.27}$$

5.3 Limitations

The proposed approaches compute the quantitative summary of various inter and intra-frame distances and angles between the joints as the feature vector. This effectively eliminates the effect of incorrect angles or distances derived from frames with occluded or noisy joint data. Simultaneously, the quantitative summarizing process may suppress the benefits of the correct frames and have a minor impact on performance.

5.4 Summary

In both approaches, we proposed novel gait event-specific features contributing to the gait cycle. The methods relied on a succinct quantitative summary of the features, drastically cutting down on the features processed by classification models. In addition, novel DL models with LSTM/GRU units were created to evaluate the temporal relationship between gait event characteristics. In the second study, the Attention unit demonstrated the efficacy of the DL model in concentrating on the most dominant features for gait recognition. The experimental evaluation of multiple datasets using various evaluation protocols revealed superior performance. The next chapter discusses single/multi view HAR system using multi-modal data and DL model.

Chapter 6

Single-/ Multi-view Human Action Recognition and Human Identification using Fusion of Data Modalities

HAR, which has many uses in surveillance and monitoring, relies heavily on RGB video data (Sun et al., 2023). Some of the earliest approaches on HAR presented in the literature rely on using conventional RGB videos (Bobick and Davis, 2001; Laptev et al., 2008). Recently, the widespread availability of low-cost depth sensors like the Microsoft Kinect (Han et al., 2013; Zhang, 2012) has resulted in multi-modal action data, which consists of RGB, depth, skeleton, and so on (Wang et al., 2020). There are benefits and drawbacks to each of these data formats. Due to their sensitivity to factors that affect the quality of RGB images, such as a complex background, illumination variation, and clothing colour, RGB-based methods make it challenging to segment the human body in certain scenes. Furthermore, the most important information for action recognition is provided by 3D data, whereas RGB data does not contain. With depth information, foreground objects in a busy background can be separated with much greater precision than with just RGB data alone. Furthermore, depth data can accurately recognize an action despite the subject's attire. As a result, research is being conducted to create a reliable action feature descriptor based on depth data for use in HAR systems. However, as reported in (Mallick et al., 2014), noise remains in depth data for many reasons. The use of skeletons to recognize actions has gained popularity due to the method's $Accuracy$ and robustness against environmental factors like lighting changes. Joint trajectories are recorded in the human skeleton. However, depth data captures the 3D structure and distance information and is thus widely used for HAR.

Due to its widespread use in various applications, there is a plethora of research on developing techniques for HAR using different data modalities. During the past several decades, researchers have examined HAR using a single modality (Ko et al., 2015; Aggarwal and Ryoo, 2011). There are several existing works on HAR, and some are based on skeleton data (Ke et al., 2017; Saggese et al., 2019), and some on depth data (Yang and Tian, 2014; Wang et al., 2015, 2016). Researchers have focused on the fusion of multiple data modalities and the transmission of information between modalities to improve the $Accuracy$ of HAR (Fan et al., 2020; Gu et al., 2020; Kamel et al., 2019).

This thesis work proposes two methods for HAR that combine skeleton and depth data modalities. Methods for efficiently representing human action using skeleton and depth data are proposed. In addition, the use of DL models in classification tasks inspired us to propose DL models for capturing spatio-temporal features from action data for accurate HAR.

Inspired by the existing works for HAR using combination of skeleton and depth data, in this chapter, we propose not only HAR system but also HI system using fusion of gait data in skeleton and depth modality. The subsequent sections provide a comprehensive description of all these works with observed results.

6.1 Skeleton and Depth Data-driven Multi-stream DL Model for Single-view HAR

In this, we propose HAR utilizing skeleton and depth data with the DL model. Following is a summary of the primary contributions of this work.

Contributions:

- Proposed a multi-stream, multi-attention DL model for HAR that explicitly learns spatio-temporal attributes of action from the depth and skeleton data.
- Conversion of the raw depth and skeleton sequence to image format, representing the action spatially and temporally.
- Proposed multiple Attention blocks to concentrate on the most distinctive body part movement while performing an action.
- Demonstrating the efficacy of the proposed system using standard evaluation protocols on two benchmark datasets with various proposed score fusions techniques.

6.1.1 Proposed Methodology

The proposed method's overall architecture for HAR is depicted in Figure 6.1. The process begins with pre-processing, followed by a CNN-based multi-stream DL model, and eventually with score fusion. This section discusses each component of the proposed model for HAR in detail.

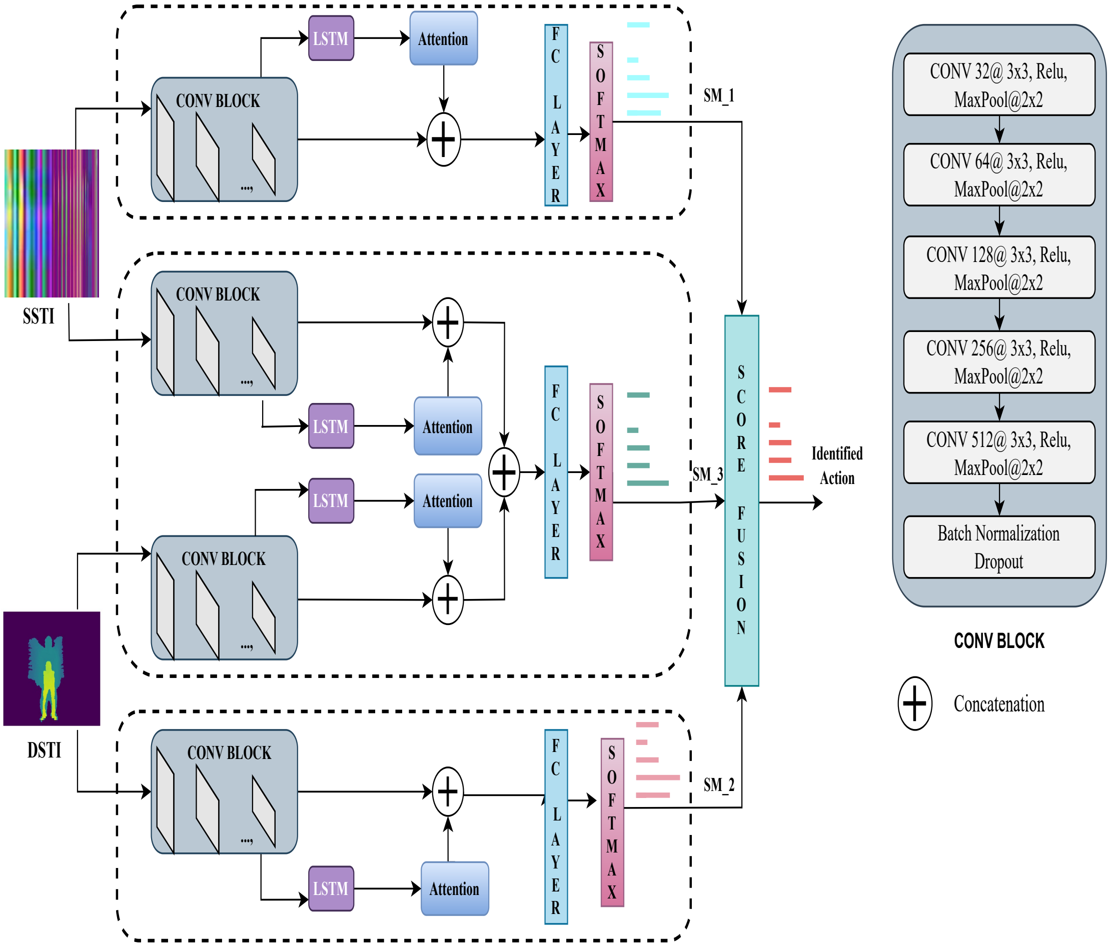

LSTM
Attention
CONV BLOCK
FC LAYER
SOFTMAX
SM_1
SSTI
SM_3
SCORE FUSION
Identified Action
DSTI
SM_2
CONV 32@ 3x3, Relu, MaxPool@2x2
CONV 64@ 3x3, Relu, MaxPool@2x2
CONV 128@ 3x3, Relu, MaxPool@2x2
CONV 256@ 3x3, Relu, MaxPool@2x2
CONV 512@ 3x3, Relu, MaxPool@2x2
Batch Normalization Dropout
CONV BLOCK
Concatenation

Figure 6.1: Proposed multi-stream multi-attention deep learning model for HAR.

6.1.1.1 Pre-processing

Each action contains a set of skeleton and depth frames, denoted by $S = \{s_1, s_2, s_3, ..., s_N\}$ and $D = \{d_1, d_2, d_3, ..., d_N\}$, where, N is the number of frames. A skeleton frame has a set of three-dimensional coordinates corresponding to the skeleton joints. The pixel value in the depth map signifies the distance between the camera's viewpoint and the scene object's surface. In the pre-processing step, the sequences of frames are transformed into a single image format that depicts the spatio-temporal features of the human action present in that sequence.

Generate Skeleton Spatio-temporal Image (SSTI)

Processing raw skeleton joints directly using a DL model with fully connected layers necessitates more computations due to the large number of data points in an action sequence. This work converts the series of skeleton frames S to an image format called Skeleton Spatio-Temporal Image (SSTI) using the three steps below.

Step1:

The method described in (Pham et al., 2019) is applied to capture the spatio-temporal postural features present in skeleton frames. All 3D coordinates are transformed into a new space by a normalization function, which is then represented as an RGB image. Consider a sequence of skeleton frames S in the space S_1. All 3D joint coordinates in space S_1 are transformed into a new space $S_1^{'}$ by a normalization function to map them into a range of 0 to 255 using Equation (6.1).

$$\begin{aligned}(x_i^t)' &= 255 \times \frac{x_i^t - min\{\varphi\}}{max\{\varphi\} - min\{\varphi\}} \\ (y_i^t)' &= 255 \times \frac{y_i^t - min\{\varphi\}}{max\{\varphi\} - min\{\varphi\}} \\ (z_i^t)' &= 255 \times \frac{z_i^t - min\{\varphi\}}{max\{\varphi\} - min\{\varphi\}}\end{aligned} \tag{6.1}$$

Where, i, t, $max\{\varphi\}$, and $min\{\varphi\}$ represent the joint number, frame number, dataset's maximum and minimum coordinate values, respectively. The new coordinate values are grouped according to the body part they correspond to: Left Hand (LH), Right Hand (RH), Trunk (T), Left Leg (LL), Right Leg (RL). Finally, all frames in the sequence are staked. Now, the new three coordinates are mapped as R, G, and B components of the image representation.

Step2:

To determine the joint transition attributes between each frame's joints the orientation features are extracted using Equation (6.2).

$$\begin{aligned}P_x^t &= \arccos \frac{(x_i)^t - (x_j)^t}{dist} \\ P_y^t &= \arccos \frac{(y_i)^t - (y_j)^t}{dist} \\ P_z^t &= \arccos \frac{(z_i)^t - (z_j)^t}{dist}\end{aligned} \tag{6.2}$$

Where, i and j represent the joint number, t denotes the frame number, and $dist$ is the magnitude of vector between the joints. Followed by this, the orientation features are normalized to the range 0 to 255 as shown in Equation (6.3).

$$\begin{aligned}(P_x^t)' &= 255 \times \frac{P_x^t - min\{\delta\}}{max\{\delta\} - min\{\delta\}}\\(P_y^t)' &= 255 \times \frac{P_y^t - min\{\delta\}}{max\{\delta\} - min\{\delta\}}\\(P_z^t)' &= 255 \times \frac{P_z^t - min\{\delta\}}{max\{\delta\} - min\{\delta\}}\end{aligned} \tag{6.3}$$

Where the $max\{\delta\}$, $\min\{\delta\}$ denote the maximum and minimum orientation features obtained using Equation (6.2). These new features are grouped and stacked similarly to $Step1$ and mapped to an image's R, G, and B components.

Step3:

The images obtained in $step1$ and $step2$ are concatenated to create a single RGB image representation of the sequence of skeleton frames named SSTI. After concatenating two distinct types of features, the entire set of values is normalized to the range 0 to 255 using the highest and lowest values from the whole feature set. The final image is resized to 224×224 pixels.

Generate Depth Spatio-Temporal Image (DSTI) representation

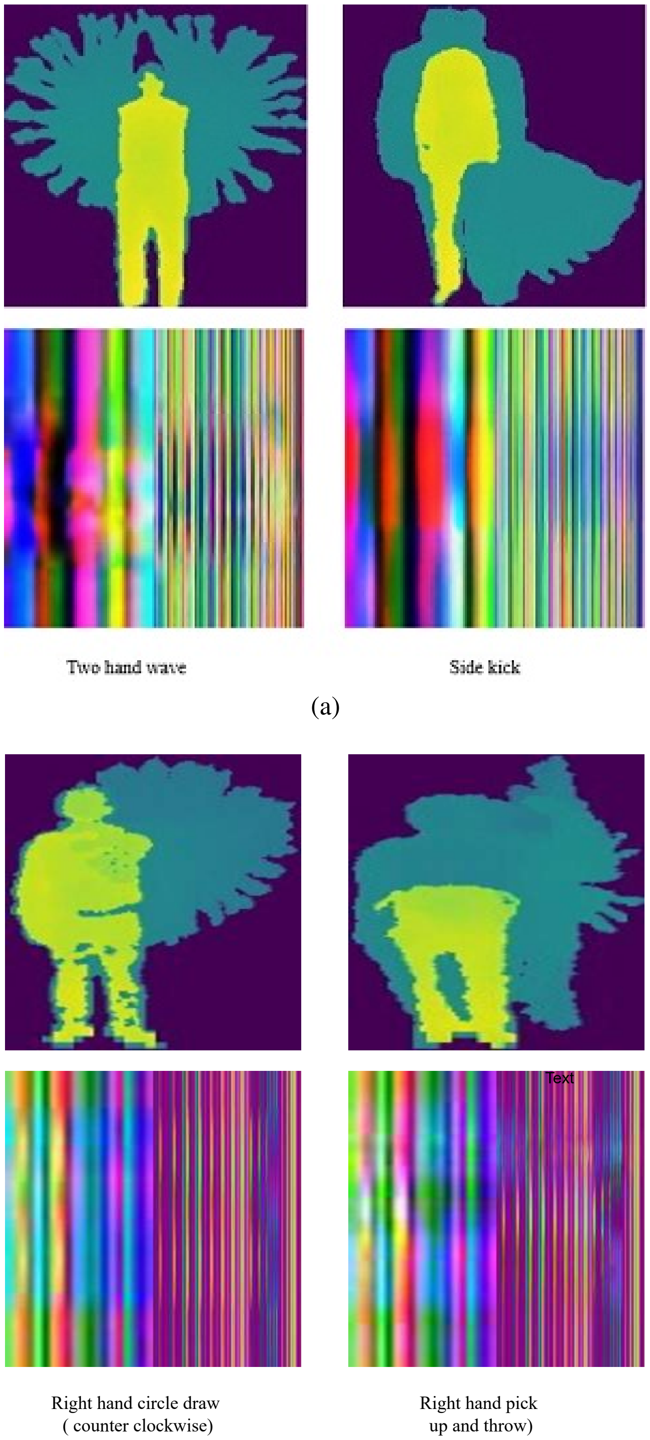
The series of depth frames D is transformed into an image as defined by Equation (6.4).

$$DSTI(i,j) = min(D\_I(i,j,t)) + max(D\_I(i,j,t)) \tag{6.4}$$

Where, i and j indicate the pixel position that ranges from 0 to the width and height of the depth frame, t denotes the frame number. Finally, images are resized to 224×224. The sample SSTI and Depth Spatio-Temporal Image (DSTI) of two actions are shown in Figure 6.2.

6.1.1.2 Attention-guided Multi-stream CNN + LSTM Model

Figure 6.1 illustrates the proposed multi-stream multi-attention DL model. Primarily three DL streams are built with CONV block, LSTM, and Attention. The SSTI,


Two hand wave
Side kick
(a)
Text
Right hand circle draw (counter clockwise)
Right hand pick up and throw)

(b)

Figure 6.2: Sample DSTI and SSTI of actions. (a) MSRAction3D. [Top: DSTI, Bottom: SSTI] (b) UTD-MHAD. [Top: DSTI, Bottom: SSTI]

and DSTI are processed by $Stream_1$, and $Stream_3$, respectively. The second stream $Stream_2$ processes both SSTI and DSTI using two sub-streams. Figure 6.1 also depicts the layers of the CNN utilized in CONV block.

Certain joints and frames in skeleton sequences were particularly distinguishable and informative for recognizing actions. For example, in the "waving hands" action, the arms' joints provide more information. These informative joints and frames constitute "crucial stages" in action. Various body parts contribute uniquely to every human action. Certain body parts play an essential role in the performance of the action. Moreover, a particular duration plays a vital role throughout the action. Consequently, additional highly informative joints/body parts, joint/body part movement, and sets of frames are especially pertinent for action in the total skeleton and depth data set. In the proposed model, CNN extracts the image representation's features. The LSTM layer captures the temporal features from output of CONV block. Multiple Attention modules were incorporated into the proposed DL model to focus on the movement of the most significant joints/body parts and frames. Each stream in the proposed DL model includes an Attention module (including sub-streams). Below the Attention module is described in detail.

Attention Module

The mechanism of (Luong et al., 2015) is followed in the Attention module of the proposed DL model. Let $\{H_t = h_1, h_2, ..., h_n\}$ be the output of the LSTM layer, where, n denotes the length of H_t. H_t is fed to the Attention module, which computes the context-specific feature vector c_t by computing a weighted sum, as defined in Equation (6.5).

$$c_t = \sum_{t=1}^{n} \Delta_t h_t \tag{6.5}$$

Where, Δ_t denotes the Attention weight calculated using a Softmax function as shown in Equation (6.6).

$$\Delta_t = \frac{e^{W_t}}{\sum_{t=1}^{n} e^{W_t}} \tag{6.6}$$

Where, the W_t is obtained as a alignment function over the output vector H_t as shown in Equation (6.7).

$$W_t = tanh(H_t) \tag{6.7}$$

The Attention scores are finally concatenated with the CNN output and further processed with a Dense layer. At the end Softmax layer generates the probability score of recognizing the action.

6.1.1.3 Score Fusion

To enhance the recognition performance of the proposed system, we proposed several score fusion techniques that exploit the capabilities of each CNN and LSTM stream. The various score fusion operations carried out on these scores are shown in Table 6.1. The Softmax scores of $Stream_1$, $Stream_2$, and $Stream_3$ are denoted as SM_1, SM_2, and SM_3, respectively. The fusion with the highest performance determines the performance of the proposed system.

Table 6.1: Score fusion operations.

Score	**Fusion Operation**
SM_1	Skeleton input $Stream_1$ Softmax
SM_2	Depth input $Stream_3$ Softmax
SM_3	Fusion input $Stream_2$ Softmax
$Fusion\_1$	$Mean(Product(SM\_1, SM\_2), Sum(SM\_1, SM\_2))$
$Fusion\_2$	$Mean(Product(SM\_1, SM\_2, SM\_3), Sum(SM\_1, SM\_2, SM\_3)$
$Fusion\_3$	$Mean(Product(SM\_2, SM\_3), Sum(SM\_2, SM\_3)$
$Fusion\_4$	$Mean(Product(SM\_1, SM\_3), Sum(SM\_1, SM\_3)$
Final	$Maximum(Fusion\_1, Fusion\_2, Fusion\_3, Fusion\_4)$

6.1.2 Experiments, Results, and Analysis

The proposed model is a multi-stream, context-aware, DL model with multiple Attention units. Each stream is equipped with a Softmax layer that calculates the classification probability. Following is a comprehensive discussion of the observation of experiments on two datasets.

6.1.2.1 Datasets

Two publicly available benchmark datasets of human actions captured with the Kinect depth sensor, namely: MSRAction3D (Li et al., 2010), and UTD Multimodal Human Action Dataset (UTD-MHAD) (Chen et al., 2015), are used to evaluate the performance of the proposed method. Chapter-4, Section 4.2.1 explains MSRAction3D in greater depth.

UTD-MHAD (Chen et al., 2015) was built using depth and inertial sensors and consists of 27 types of human actions performed by eight individuals. The Kinect depth sensor was mounted on a tripod approximately 3 meters in front of the subject to capture images of the subject's entire body during data creation. The actions are performed four times by each individual. After removing corrupted sequences, the dataset consists of 861 samples. Due to the following factors, this dataset demonstrates significant intra-class variation: 1) The subject carried out the same action at varying speeds across trials. 2) The heights of the subjects are different. 3) The same action was repeated in a natural manner, thereby making each trial unique. Four modalities of data were provided, including RGB video, depth video, skeleton joints, and an inertial sensor signal. In this work, we utilized depth and skeleton data of human action videos.

6.1.2.2 Evaluation Protocol and Metrics

The performance of the proposed HAR system is evaluated using $Accuracy$. To ensure a fair comparison, the MSRAction3D dataset was evaluated using the well-established assessment technique known as the Cross-Subject ($C-S$) protocol. Here, subjects with odd numbers (1, 3, 5, 7, and 9) are used for training, whereas subjects with even numbers (2, 4, 6, 8, and 10) are used for testing. In addition, the dataset is divided into three groups: AS1, AS2, and AS3, according to the baseline (Li et al., 2010). On UTD-MHAD, the assessment protocol is identical to the previous one in that it entails training with odd subjects (1, 3, 5, 7) and testing with even subjects (2, 4, 6, 8).

6.1.2.3 Experiments and Results

A series of experiments were conducted with SSTI and DSTI independently, as well as in conjunction with the different streams of the proposed multi-stream DL model. Table 6.2 summarises the $Accuracy$ of the proposed system in various experimental setups utilizing a $C-S$ evaluation method on the MSRAction3D and UTD-MHAD datasets. Table 6.2 demonstrates that the proposed model (CNN + LSTM + Attention) outperformed the CNN model without LSTM and Attention.

LSTM networks are exceptionally effective at capturing temporal characteristics. We used LSTM in conjunction with CNN to capture these features, as human action involves constantly changing features over time. Considering all of the temporal features, the movement of a specific body part at a particular duration best describes the action performed. Thus, Attention is employed to focus on a particular set of characteristics. Combining multiple Attentions with LSTM demonstrated their effectiveness in

Table 6.2: Recognition $Accuracy$ (in %) of the proposed system on MSRAction3D and UTD-MHAD.

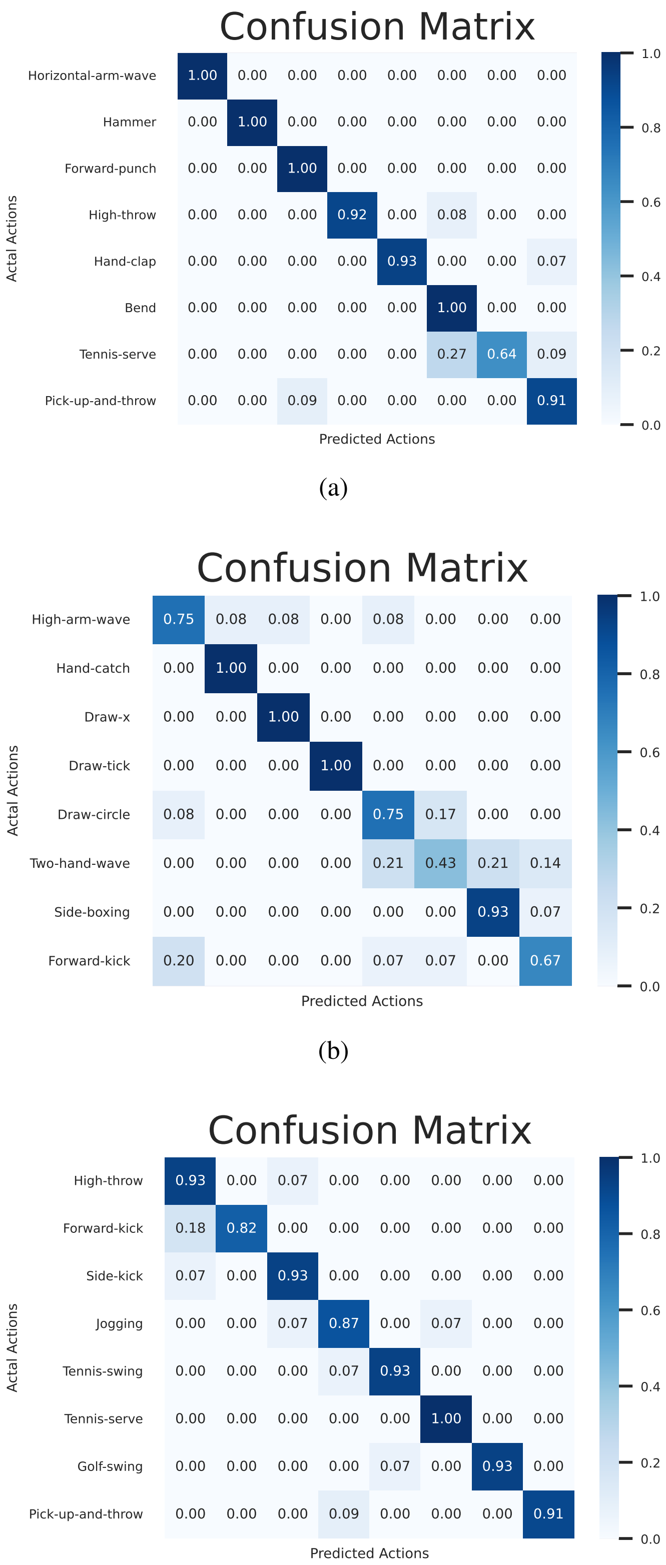
Model	**Fusion Score**	**Dataset**			
		MSRAction3D			**UTD-MHAD**
		AS1	**AS2**	**AS3**	
CNN+LSTM +Attention	SM_1	81.1	82.3	85.7	87.2
	SM_2	90.6	74.3	90.2	58
	SM_3	89.7	78.8	91.1	87.9
	$Fusion\_1$	88.7	81.4	90.1	78.6
	$Fusion\_2$	91.5	82.3	89.3	90.0
	$Fusion\_3$	93.4	78.8	93.8	77.0
	$Fusion\_4$	92.3	81.4	92.0	90.7
	Final	**93.4**	**82.3**	**93.8**	**90.7**
CNN	SM_1	76.4	81.4	83.9	86.4
	SM_2	88.7	73.5	89.3	54.4
	SM_3	86.8	80.5	90.1	85.8
	$Fusion\_1$	87.7	81.4	91.1	77.7
	$Fusion\_2$	88.8	81.5	89.3	85.8
	$Fusion\_3$	87.7	81.4	92.0	76.0
	$Fusion\_4$	89.6	80.5	91.1	88.6
	Final	89.6	81.45	91.1	88.6

capturing the most important spatio-temporal data necessary for accurately recognizing human actions.

The confusion matrices for MSRAction3D and UTD-MHAD are depicted in Figures 6.3 and 6.4, respectively. The confusion matrices of AS1, AS2, and AS3 are depicted in Figures 6.3 (a), 6.3 (b), and 6.3 (c), respectively. The majority of actions are perfectly recognized in both datasets.

6.1.2.4 Comparison with Existing Works

The proposed system demonstrated superior recognition performance on these two datasets compared to various prior works. In Tables 6.3 and 6.4, the $Cross - Subject$ protocol comparisons between the proposed HAR system and various base-line works on datasets: MSRAction3D and UTD-MHAD are detailed.



Horizontal-arm-wave	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Hammer	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Forward-punch	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
High-throw	0.00	0.00	0.00	0.92	0.00	0.08	0.00	0.00
Hand-clap	0.00	0.00	0.00	0.00	0.93	0.00	0.00	0.07
Bend	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Tennis-serve	0.00	0.00	0.00	0.00	0.00	0.27	0.64	0.09
Pick-up-and-throw	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.91

(a)

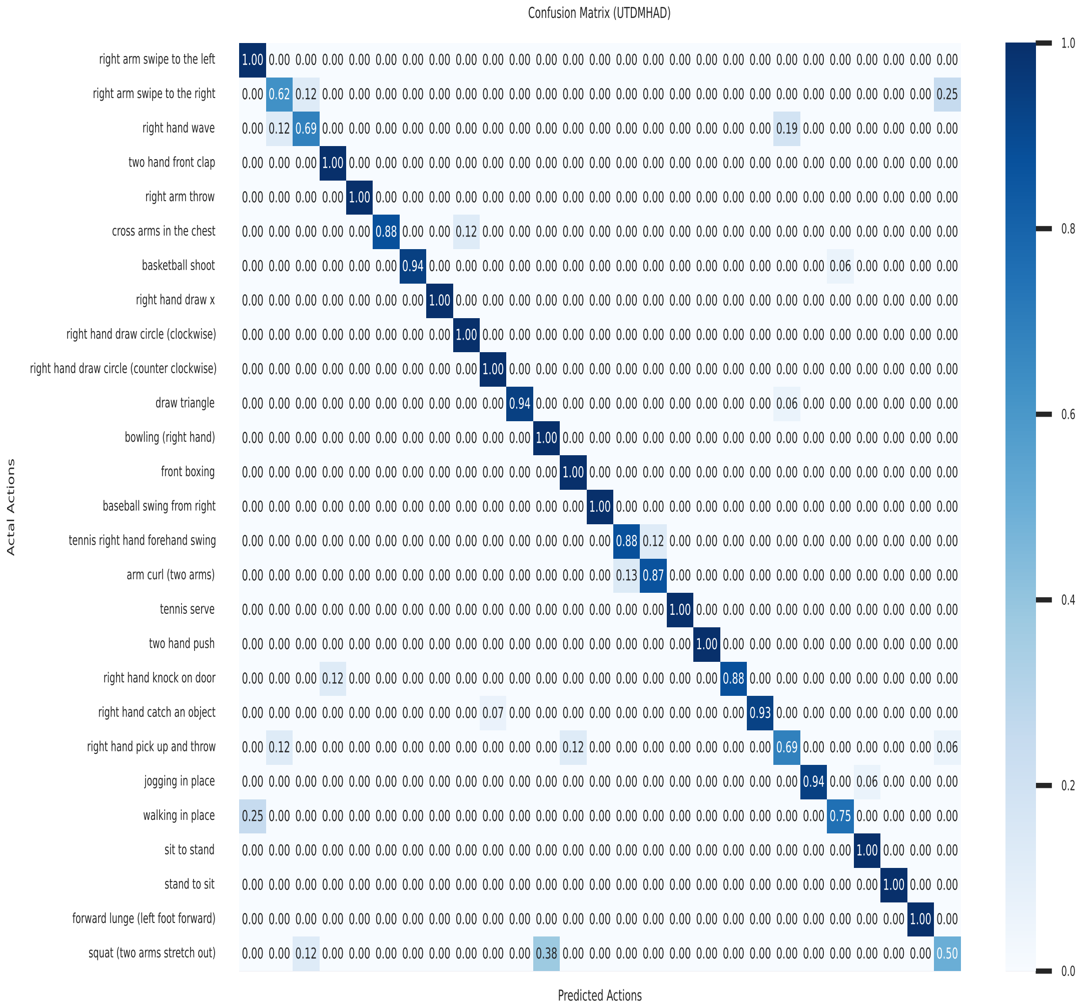
High-arm-wave	0.75	0.08	0.08	0.00	0.08	0.00	0.00	0.00
Hand-catch	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
Draw-x	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
Draw-tick	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
Draw-circle	0.08	0.00	0.00	0.00	0.75	0.17	0.00	0.00
Two-hand-wave	0.00	0.00	0.00	0.00	0.21	0.43	0.21	0.14
Side-boxing	0.00	0.00	0.00	0.00	0.00	0.00	0.93	0.07
Forward-kick	0.20	0.00	0.00	0.00	0.07	0.07	0.00	0.67

(b)

High-throw	0.93	0.00	0.07	0.00	0.00	0.00	0.00	0.00
Forward-kick	0.18	0.82	0.00	0.00	0.00	0.00	0.00	0.00
Side-kick	0.07	0.00	0.93	0.00	0.00	0.00	0.00	0.00
Jogging	0.00	0.00	0.07	0.87	0.00	0.07	0.00	0.00
Tennis-swing	0.00	0.00	0.00	0.07	0.93	0.00	0.00	0.00
Tennis-serve	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Golf-swing	0.00	0.00	0.00	0.00	0.07	0.00	0.93	0.00
Pick-up-and-throw	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.91

(c)

Figure 6.3: Confusion matrix obtained for MSRAction3D. (a) AS1. (b) AS2. (c) AS3.



right arm swipe to the left	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
right arm swipe to the right	0.00	0.62	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25
right hand wave	0.00	0.12	0.69	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.00	0.00	0.00	0.00	0.00	0.00
two hand front clap	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
right arm throw	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
cross arms in the chest	0.00	0.00	0.00	0.00	0.00	0.88	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
basketball shoot	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00
right hand draw x	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
right hand draw circle (clockwise)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
right hand draw circle (counter clockwise)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
draw triangle	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00
bowling (right hand)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
front boxing	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
baseball swing from right	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
tennis right hand forehand swing	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.88	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
arm curl (two arms)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.87	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
tennis serve	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
two hand push	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
right hand knock on door	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.88	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
right hand catch an object	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.93	0.00	0.00	0.00	0.00	0.00	0.00	0.00
right hand pick up and throw	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.69	0.00	0.00	0.00	0.00	0.00	0.06
jogging in place	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.00	0.06	0.00	0.00	0.00
walking in place	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.75	0.00	0.00	0.00	0.00
sit to stand	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
stand to sit	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
forward lunge (left foot forward)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
squat (two arms stretch out)	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.38	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50

Figure 6.4: Confusion matrix obtained for UTD-MHAD.

Table 6.3: Proposed system comparison with existing methods on MSRAction3D.

Methodology	**AS1**	**AS2**	**AS3**	**Overall**
Bag_of_3D (Li et al., 2010)	72.9	71.9	79.2	74.7
HODJ (Xia et al., 2012)	87.98	85.48	63.46	78.97
EigenP (Yang and Tian, 2012)	74.5	76.1	96.4	82.3
FVSQ (Evangelidis et al., 2014)	88.39	**86.61**	**94.59**	89.8
Proposed Method	**93.4**	82.3	93.8	**89.83**

Table 6.4: Proposed system comparison with existing methods on UTD-MHAD.

Method	**Accuracy**
Kinect and Inertial (Chen et al., 2015)	79.10
JDMS (Li et al., 2017)	88.10
SDSR (Annadani et al., 2016)	86.12
SOS (Hou et al., 2018)	86.97
JTM (Wang et al., 2018)	87.90
DCNN (Kamel et al., 2019)	88.14
DSIEMM (Yang et al., 2020)	88.37
Proposed Method	**90.7**

6.2 Multi-stream Attention-guided Deep Networks with Skeleton and Depth Data from Overlapping Sub-actions for Single-/Multi-view HAR

Inspired by the advantages of using multi-modal data, this work proposes a DL-based method with Attention units for HAR employing skeleton and depth data. Furthermore, the studies of human actions have shown that each action is actually made up of a number of smaller sub-actions (Liang et al., 2020). The action of "throwing a ball", for instance, consists of a series of smaller actions, such as "move hand toward ball," "pick ball," "hold ball in hand and move hand up," "move hand to front," and so on. The number of sub-actions depends on the complexity of the action. So, we represented the actions by exploring a range of sub-actions' worth of data, based on the skeleton and depth information. Furthermore, we employed a number of Attention blocks to focus on crucial spatio-temporal features for HAR. The key contributions of the proposed work are given below.

Contributions:

- A method for exploiting the features of overlapping sub-actions is proposed.
- Developed a novel depth-based action descriptor that combines the sub-actions in action video, thereby reducing the number of features to be processed.
- Extraction of summarised features from overlapping sub-actions in a sequence of

skeleton frames by partitioning the human body into five regions, which substantially reduced the number of features to be processed.

- Proposed a multi-stream Attention-based deep neural network model for efficiently learning the spatio-temporal characteristics from multi-modal data.

6.2.1 Proposed Methodology

This study proposes two distinct pre-processing procedures for action videos in two data modalities. As a human action is a series of overlapping sub-actions, we attempted to describe it using sub-action features. Moreover, a multi-stream DL model with Attention is proposed to learn the spatio-temporal features of these overlapping sub-actions. We combine features from two modalities and scores from multiple deep-learning streams to achieve optimal performance. Figure 6.5 depicts the workflow for the proposed work. More in-depth explanation of the proposed work is given below.

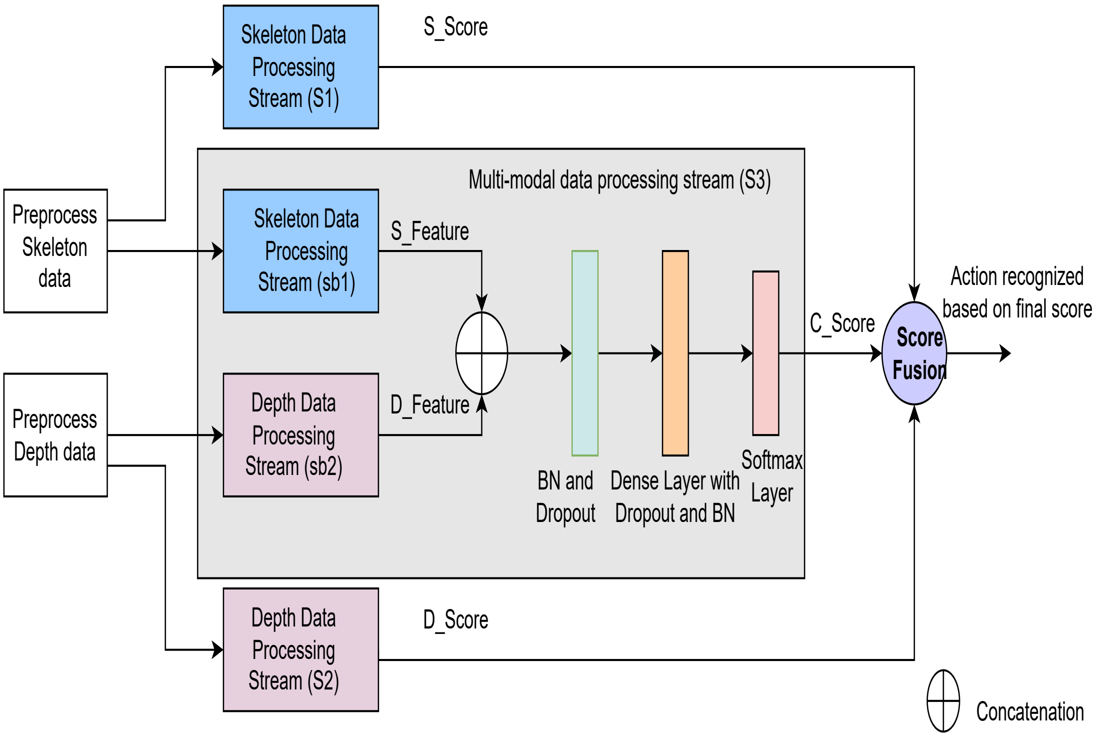

Skeleton Data Processing Stream (S1)
S_Score
Multi-modal data processing stream (S3)
Preprocess Skeleton data
Skeleton Data Processing Stream (sb1)
S_Feature
Action recognized based on final score
C_Score
Score Fusion
Preprocess Depth data
Depth Data Processing Stream (sb2)
D_Feature
Softmax Layer
BN and Dropout
Dense Layer with Dropout and BN
Depth Data Processing Stream (S2)
D_Score
Concatenation

Figure 6.5: Work flow architecture of the proposed HAR system.

6.2.1.1 Pre-processing

To effectively represent action data, we pre-process a series of depth images and 3D skeleton data of human movements during an action. Consider a series of skeleton frames $F = \{f_1, f_2, f_3..., f_N\}$, and depth frames $D = \{d_1, d_2, d_3..., d_N\}$ from a given N-frame human action video.

Skeleton Data

We conduct a quantitative summary of skeleton data to reduce the effect of noisy joints on the overall performance of the proposed system. In turn, it is an attempt to reduce the number of features and computations. Since the proposed method considers the multi-view human action dataset, the data from different view points are merged into a single, view-independent global pool by transforming the skeleton data's 3D coordinates into a global coordinate system centered on $(0, 0, 0)$. With this transformation, as shown in Equation (6.8), the hip center (hx, hy, hz) becomes the origin with coordinate values $(0, 0, 0)$. Similar methods are employed for the single-view data set.

$$\begin{bmatrix} x_i^{'} \\ y_i^{'} \\ z_i^{'} \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & -h_x \\ 0 & 1 & 0 & -h_y \\ 0 & 0 & 1 & -ht_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ z_i \\ 1 \end{bmatrix}, \tag{6.8}$$

Where, (x_i, y_i, z_i) represents original 3D coordinate values of i^{ih} joint, and $(x_i^{'}, y_i^{'}, z_i^{'})$ denotes the corresponding 3D coordinate values in global pool. $(-h_x, -h_y, -h_z)$ is the translation vector used to translate hip center to $(0, 0, 0)$.

To effectively exploit the sub-actions, the skeleton frames of an action video are grouped into overlapping sets of ten sub-actions, as shown in Figure 6.6.

As given in Figure 6.7, the skeleton joints are organized into five regions: G1, G2, G3, G4, and G5. Considering the skeleton frames in each of the ten sub-actions, the average coordinate values for each joint are computed. Consequently, the number of joint coordinates that must be analyzed for action recognition is drastically decreased. Using the proposed region-based residual LSTM models, the average 3D coordinate values of a sequence of sub-actions are learned for action recognition.

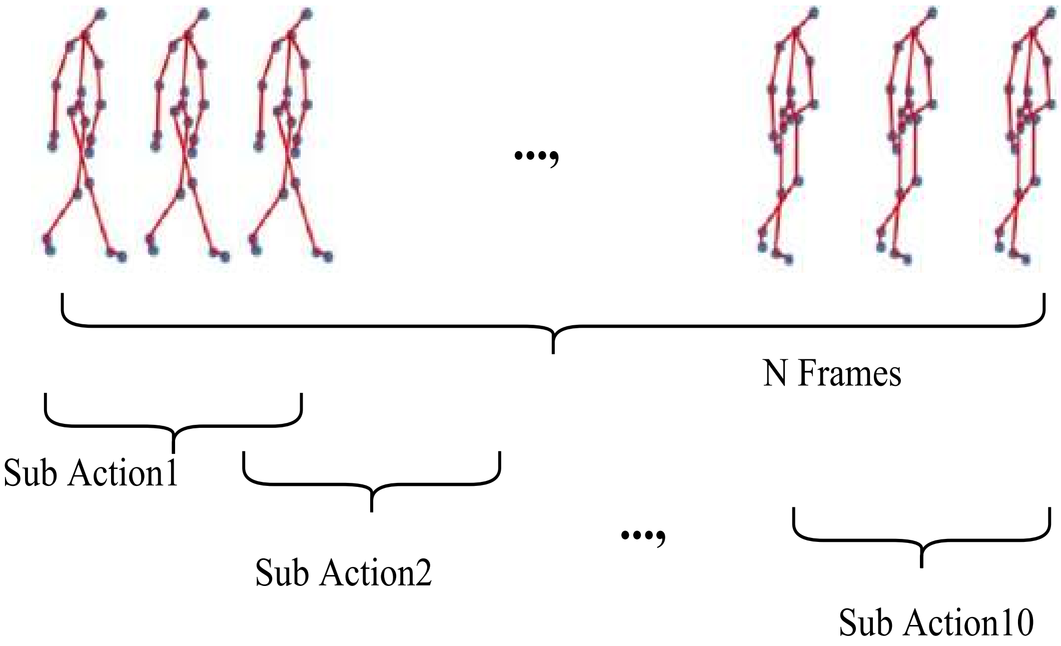
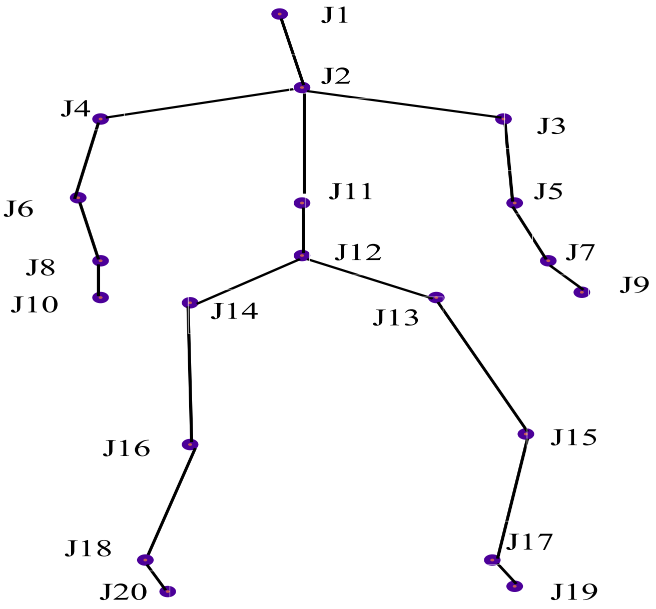


Figure 6.6: Arranging sequence of frames into overlapping sub-actions.



G1: {J1, J2, J4, J6, J8, J10}
G2: {J1, J2, J3, J5, J7, J9}
G3: {J1, J2, J11, J12, J13, J14}
G4: {J11, J12, J14, J16, J18, J20}
G5: {J11, J12, J13, J15, J17, J19}

Figure 6.7: Human skeleton joints considered and five regions

Depth Data

The depth frame sequence D is organized into four distinct, overlapping sub-actions. To effectively capture the spatial and temporal information from a sequence of depth frames, we downsample these sequences into a single image, a Multiple Sub-action Enhanced Depth Motion Map (MS-EDMM). The procedure for generating MS-EDMM from four sub-actions: DS_1, DS_2, DS_3, and DS_4, is outlined below.

At first, for each sub-action Enhanced Depth Motion Map (EDMM) is generated as follows: Let depth frames in sub-action DM_i be $\{d_0, d_1, d_2, d_3, ...d_t\}$. To begin with, the Depth Motion Energy (DME) is estimated for each sub-action as the absolute difference between two consecutive depth maps, in contrast to (Yang et al., 2012). Then, Sub-action Depth Motion Map (SDMM) is computed by adding together DME images encompassing the entire depth maps of a sub-action, as shown in Equation (6.9).

$$SDMM = \sum_{i=1}^{t} |d_i - d_{i-1}| \tag{6.9}$$

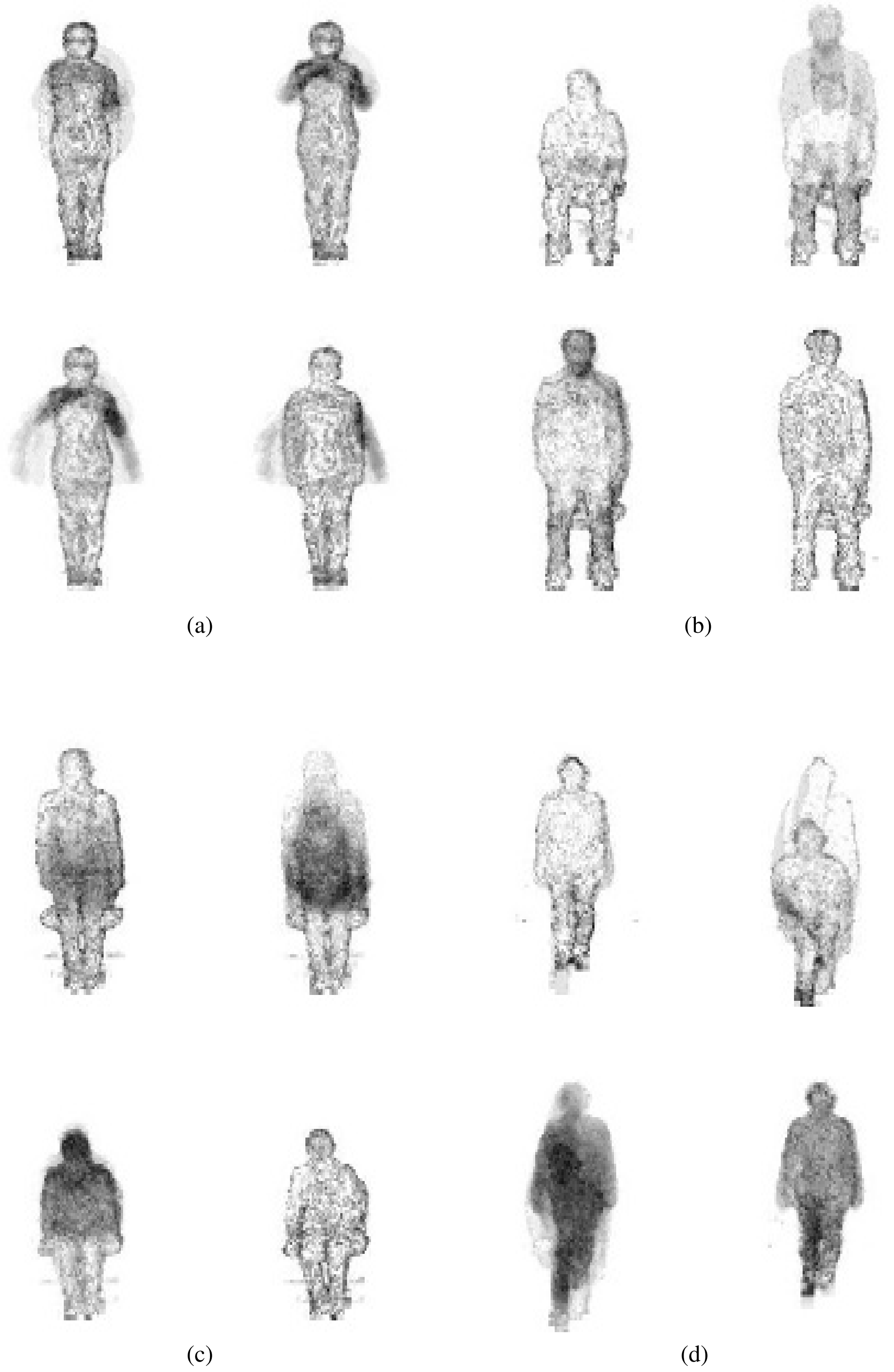
Where, $|d_i - d_{i-1}|$ denotes the DME image produced by two consecutive depth maps d_i and d_{i-1}, and i indicates the frame number. The SDMM is further improved by applying Equation (6.10) to obtain EDMM.

$$EDMM = 255 - \left(\frac{SDMM - min(SDMM)}{max(SDMM) - min(SDMM)} * 255 \right) \tag{6.10}$$

Concatenation of EDMM images from four sub-actions, namely: DS_1, DS_2, DS_3, and DS_4 produced MS-EDMM of the given action. The sample MS-EDMM generated for various actions taken from UTD-MHAD, and NTU RGB+D dataset are depicted in Figures 6.8 and 6.9, respectively.

6.2.1.2 Multi-stream Attention-based Deep Neural Network Model

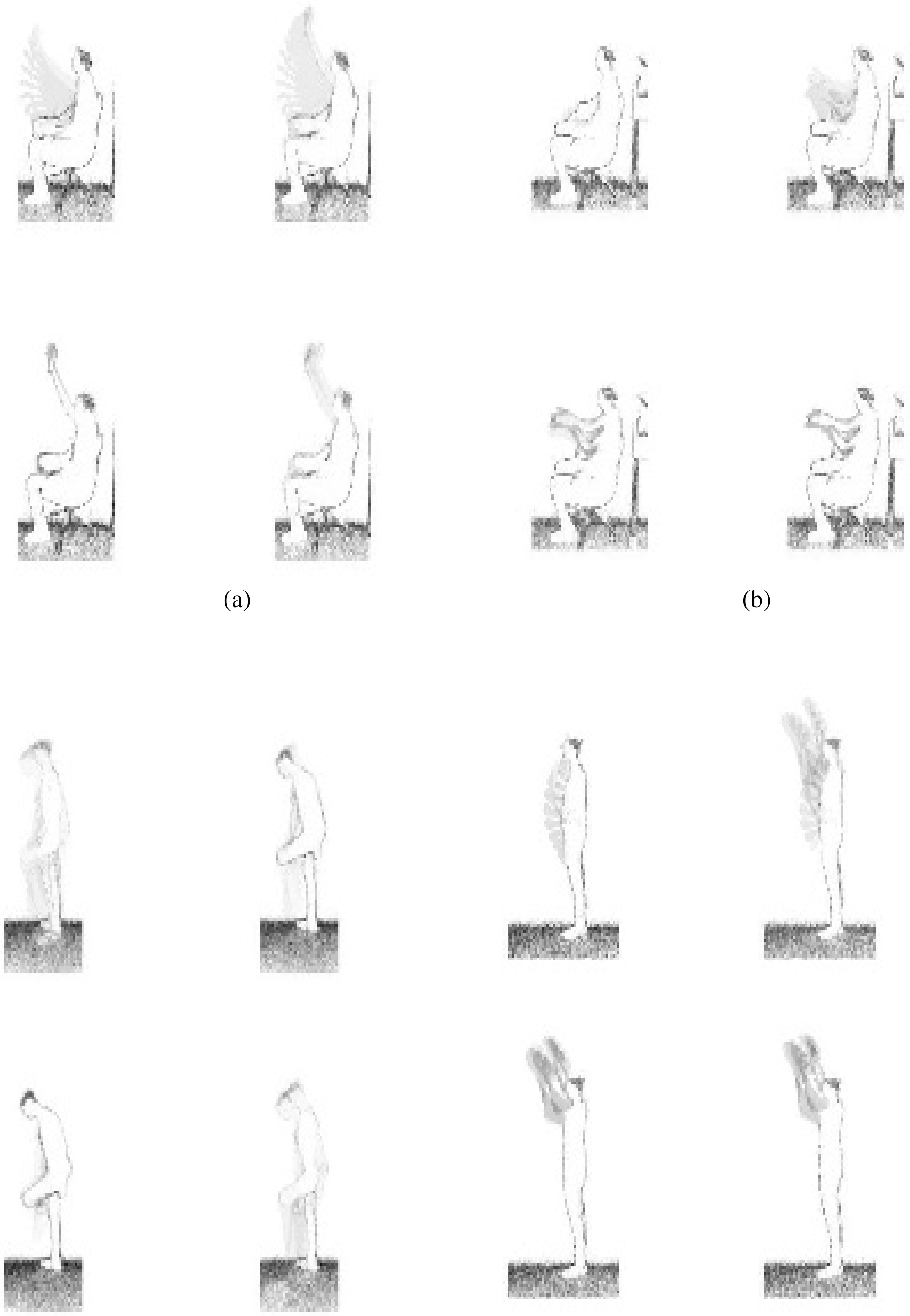
This work proposes a multi-stream DL model with CNN, residual LSTM, and Attention units to train pre-processed action video skeleton features and depth descriptors. The proposed DL model consists of three streams labeled S1, S2, and S3. Skeleton data



(a) (b)

(c) (d)

Figure 6.8: Sample MS-EDMM images generated for actions from UTD-MHAD dataset. (a) Boxing. (b) Sit to stand. (c) Stand to sit. (d) Lunge.



(a) (b) (c) (d)

Figure 6.9: Sample MS-EDMM images generated for actions from NTU RGB+D dataset: (a) Hand waving. (b) Rub two hands. (c) Take off a shoe. (d) Cheer up

from five different body parts are used to train five separate region-based LSTM models, each with a temporal Attention unit, in Stream S1. The spatio-temporal features from MS-EDMM are processed by CNN+Spatial Attention and residual LSTM with temporal Attention in Stream S2.

The features from multiple data streams are combined with a stream S3, which consists of two sub-streams, sb1 and sb2. Where, sb1 and sb2 are the replicas of S1 and S2, respectively, except the classification layer. The proposed streams S1 and S2 are depicted in Figures 6.10 and 6.11, respectively.

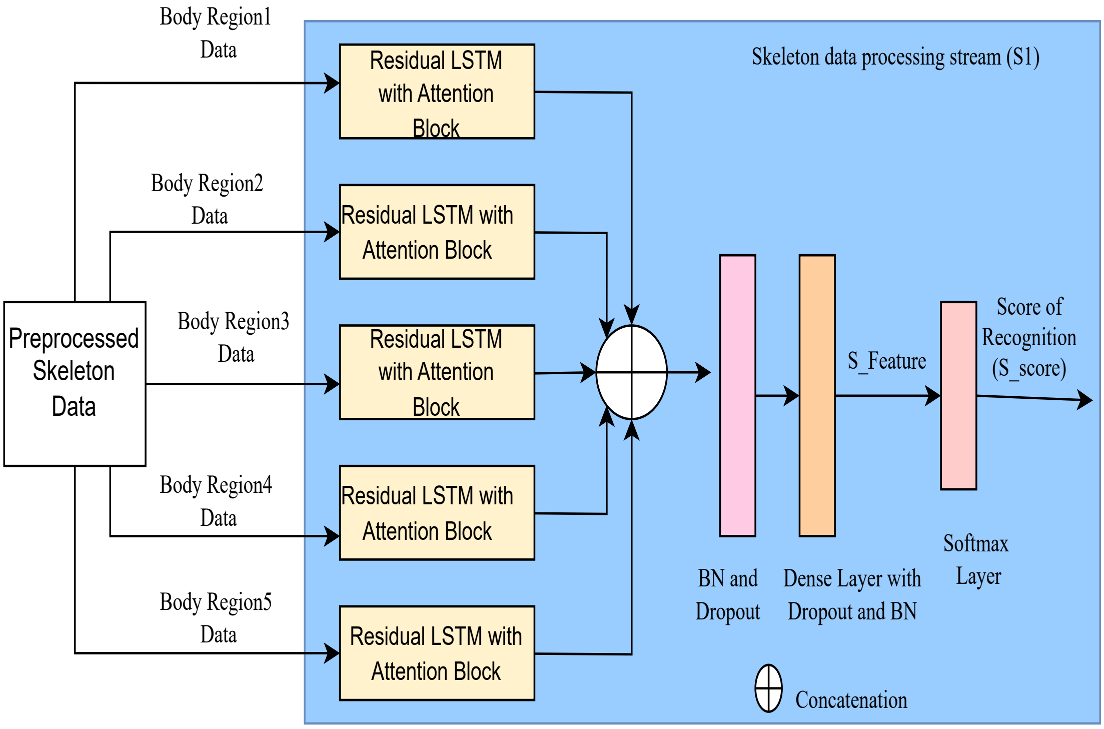

Body Region1 Data
Body Region2 Data
Body Region3 Data
Body Region4 Data
Body Region5 Data
Preprocessed Skeleton Data
Residual LSTM with Attention Block
Residual LSTM with Attention Block
Residual LSTM with Attention Block
Residual LSTM with Attention Block
Residual LSTM with Attention Block
Skeleton data processing stream (S1)
BN and Dropout
Dense Layer with Dropout and BN
S_Feature
Softmax Layer
Score of Recognition (S_score)
Concatenation

Figure 6.10: Skeleton data processing stream (S1).

Skeleton Data Processing Stream (S1)

Five regional residual LSTM with temporal Attention are used to process the pre-processed skeleton data, which consists of the mean coordinate values for each skeleton joint in ten separate sub-actions from five different regions: G1, G2, G3, G4, and G5. The spatio-temporal features from these five LSTM blocks are combined and further processed by a Dense layer with 'ReLU' activation function and produces the trained feature set called 'S_Feature'. Finally, the Softmax layer generates the recognition score

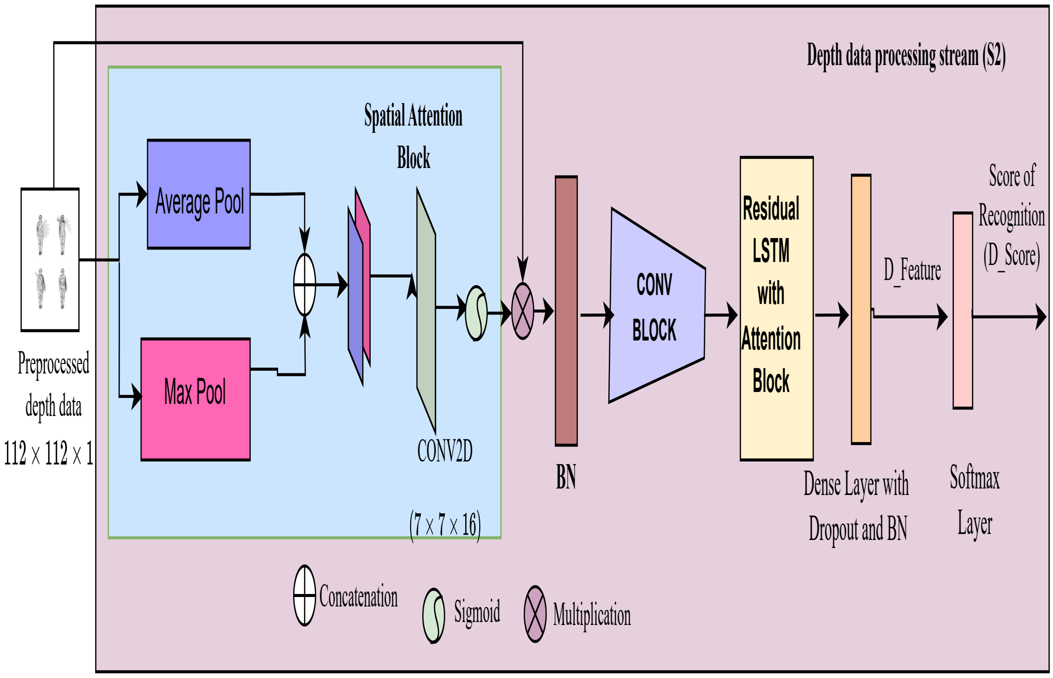


Figure 6.11: Depth data processing stream (S2).

'S_score'. In the S1 stream, the residual LSTM with Attention block consists of three residual LSTM layers with 18 units and a temporal Attention block to focus on more dominant features for action recognition as shown in Figure 6.12. Finally, The Softmax layer computes the score of recognition S_Score using the skeleton data.

It has been demonstrated that LSTM (Hochreiter and Schmidhuber, 1997) effectively addresses the vanishing gradient issue of the RNN. The proposed model incorporates a residual connection between the LSTM layers, which was inspired by the idea of attaching a skip connection among adjacent layers, which has shown promising results for training DL models (Wu et al., 2016) (He et al., 2016). The detailed view is shown in Figure 6.12. Where, let x_t^i, h_t^i be the input and hidden state output, respectively, of i^{th} LSTM layer at time t. Then the input to $(i+1)^{th}$ layer will be the element-wise sum of input to and output from the i^{th} layer as defined in Equation (6.11).

$$x_t^{(i+1)} = x_t^i \oplus h_t^i \tag{6.11}$$

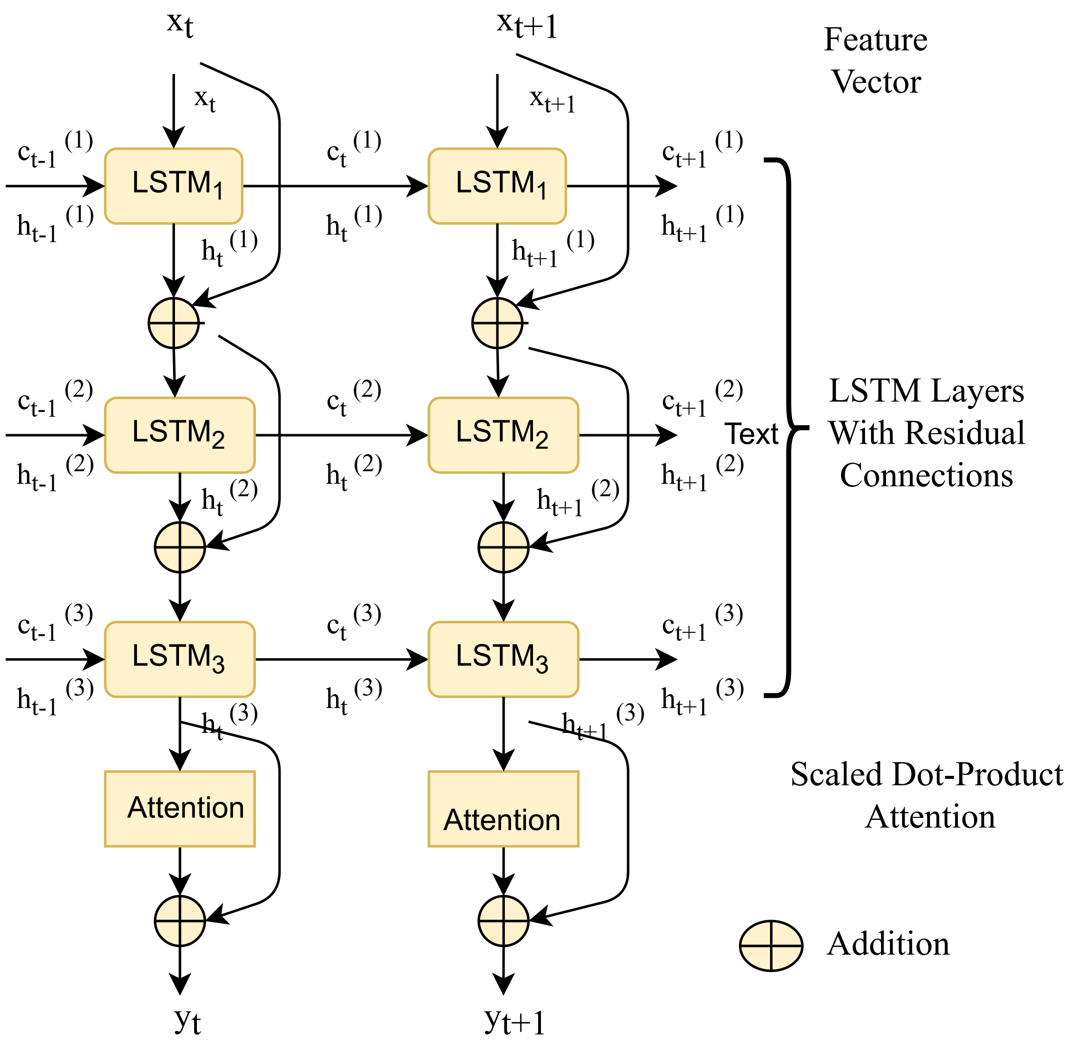


Figure 6.12: Proposed residual LSTM with Attention block.

Attention Unit (Temporal Attention)

The Attention unit improves the performance of the system by focusing on the most important context-specific features in the data streams. We employed a variation of the Self-Attention algorithm called "Scaled Dot-Product Attention" (Vaswani et al., 2017) as temporal Attention unit to produce the context-specific features from the features generated by residual LSTM layers. The technical concept of "Scaled Dot-Product Attention" is already explained in Chapter-5, Section 5.2.1.3 (Page No. 111).

Depth Data Processing Stream (S2)

The proposed depth data processing stream (S2) is used to learn MS-EDMM images generated from depth images. Figure 6.11 shows the details of S2. As MS-EDMM is a combination of four images, the Spatial Attention (Woo et al., 2018) is applied to focus on more important features in the image to generate the Spatial Attention Map S_AM for the image using Equation (6.12). Then, the original image is multiplied with the S_AM to focus on important regions in the image. The generated output matrix is further processed using the proposed CONV block. The features from this CONV block are trained using the residual LSTM with Attention block to focus on more significant spatio-temporal features for action recognition.

$$S_AM = \sigma\left(C^{7\times7\times16}[AvgPool(I); MaxPool(I)]\right) \tag{6.12}$$

Where, I, C represent the MS-EDMM image, and CNN layer.

The detailed view of the CONV block is shown in Figure 6.13. Here, there are five CNN layers with 'ReLU' activation. After each CNN layer, Batch Normalization (BN) and Dropout of 0.2 is used to avoid over-fitting and stabilize the network during training. The Residual LSTM with Attention block is the same as the one used in stream S1, except the number of units used here is 9. Following the residual LSTM with Attention block, a Dense layer is used before the Softmax layer. The Dense layer generates the final feature output $D\_Feature$. The number of units in Dense layers is fixed as four times the number of labels in the dataset. The Softmax layer produces the score of recognition D_Score using the depth data.

Multi-modal Data Processing Stream (S3)

The proposed DL model has a third stream that uses both skeleton and depth information. Hence, it is comprised of two sub-streams: $sb1$ and $sb2$, which are replicas of streams: S1 and S2 (excluding Softmax). In the end, feature fusion of $S\_Feature$, and $D\_Feature$ is performed. The combined feature is learned with a Dense layer, and finally, the Softmax layer computes the score of action recognition C_Score.

6.2.1.3 Score Fusion

In the proposed DL model, the score fusion of multiple streams is performed to compensate for the shortcomings of various data modalities and actually focus on their strengths. S_Score, D_Score, and C_Score represent the Softmax scores: $S1$, $S2$, and $S3$, respectively. The various fusion operations carried out on these are shown in Table

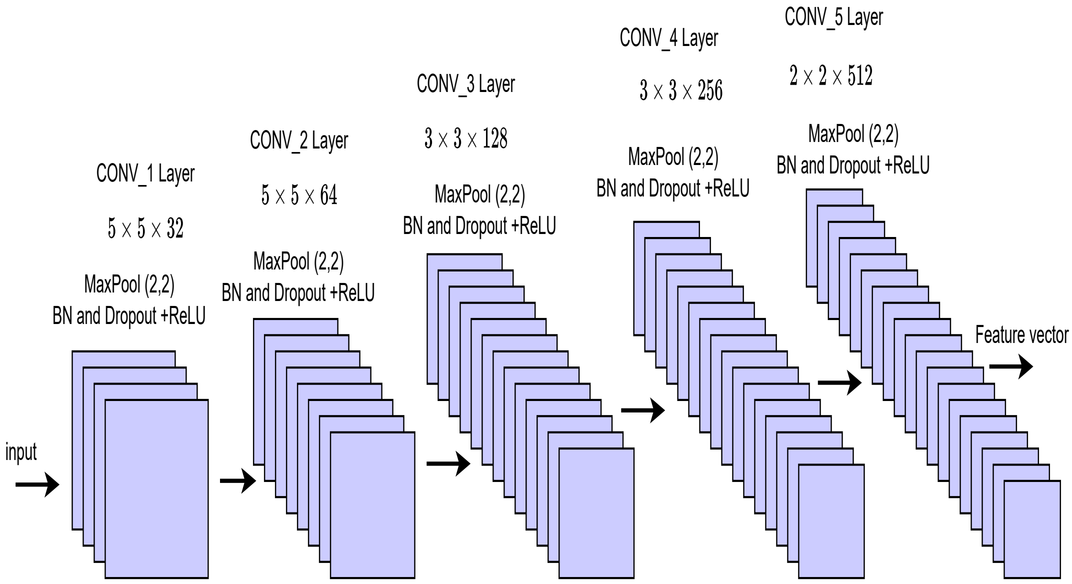


Figure 6.13: CONV block in proposed depth data processing stream.

6.5. The optimal performance fusion operation is considered as the proposed system's performance.

Table 6.5: Score fusion operations.

Score	**Fusion Operation**
S_Score	$S1$ Softmax
D_Score	$S2$ Softmax
C_Score	$S3$ Softmax
$Fusion\_1$	$Product(S\_Score, D\_Score)$
$Fusion\_2$	$Sum(S\_Score, D\_Score)$
$Fusion\_3$	$Product(S\_Score, C\_Score)$
$Fusion\_4$	$Sum(S\_Score, C\_Score)$
$Fusion\_5$	$Product(D\_Score, C\_Score)$
$Fusion\_6$	$Sum(D\_Score, C\_Score)$
$Fusion\_7$	$Product(S\_Score, D\_Score, C\_Score)$
$Fusion\_8$	$Sum(S\_Score, D\_Score, C\_Score)$

6.2.2 Experiments, Results, and Analysis

The performance of the proposed HAR system is evaluated using benchmark evaluation protocols on the small-scale (single-view) dataset: UTD-MHAD (Chen et al., 2015) and large-scale (multi-view) dataset: NTU RGB+D (Shahroudy et al., 2016).

6.2.2.1 Dataset

The details about the single-view dataset is given in Section 6.1.2.1 (Page No. 134). The NTU RGB+D dataset (Shahroudy et al., 2016) is a large-scale RGB+D dataset collected for HAR using three Microsoft Kinect v2 depth sensors. This provides data in different modalities including RGB frames, depth sequences, skeleton data, and infrared frames. It constitutes 56880 action samples and four million frames. Forty individuals in the dataset perform sixty distinct human actions. During data capture, three depth sensors were simultaneously deployed at the same height from three different horizontal angles: -45°, 0°, and +45°. The subject performed each action twice, facing either the left or right sensor. Adjustments were made to the height of the sensors and their distances from the subject to obtain additional viewpoint variations. Therefore, there are 80 distinct viewpoints. The dataset includes several types of actions, including daily actions, medical conditions, and mutual actions. As data is captured using Kinect v2, it provides information about 25 joints per skeleton. The proposed work uses 20 joints as given by Kinect v1 and discards other joints.

6.2.2.2 Evaluation Protocols and Metrics

The proposed system is evaluated on UTD-MHAD using $C-S$ protocol as explained in Section 6.1.2.2 (Page No. 135). The NTU RGB+D has two standard evaluation protocols, namely: $C-S$ and Cross-View ($C-V$), as described in the (Shahroudy et al., 2016), to assess the performance. In the $C-S$ protocol, samples from 20 subjects are used for training the network, while samples from the remaining 20 subjects are used for testing. The subjects used for training are: 1, 2, 4, 5, 8, 9, 13, 14, 15, 16, 17, 18, 19, 25, 27, 28, 31, 34, 35, and 38. Whereas, in $C-V$ protocol, the samples from Cameras 2 and 3 are used for training the network and tested using the samples from Camera 1. The performance of proposed system is measured using $Accuracy$. Also, $F-Score$ is utilized to compare the performance among the set of action labels.

6.2.2.3 Experiments and Results

A range of experiments were conducted on UTD-MHAD and NTU RGB+D datasets to test the efficiency of proposed HAR system using skeleton and depth data.

Training Parameters

The proposed multi-stream neural network for HAR is trained using Categorical cross-entropy loss function and Adam optimizer (Kingma and Ba, 2014) with initial learning rate set to 0.005. The learning rate is decreased by the factor of 0.1, and 0.5 if no improvement found in validation loss for 10 epochs in NTU-RGB+D, and UTD-MHAD, respectively. Also we stopped the training if there is no improvement found in validation $Accuracy$ for 75, and 25 epochs in UTD-MHAD, and NTU RGB+D, respectively. Also, based on experimental observations the dropout for last layer is increased to 0.5 while training the UTD-MHAD actions. In addition, l2 regularization of 0.0001 and 0.01 for CNN layers in CONV net on NTU RGB+D and UTD-MHAD, respectively.

Experiments

Several experiments were conducted to determine the significance of the number of layers in residual LSTM networks in action recognition by varying the number of layers in residual LSTM model.

Table 6.14 reports the experimental results in terms of $Accuracy$ on UTD-MHAD dataset using $C-S$ protocol. Tables 6.7 and 6.8 shows the $Accuracy$ obtained on three streams, and on different score fusion operations on NTU RGB+D dataset using $C-S$ and $C-V$ evaluation protocols, respectively. As part of the ablation study, the experiments are conducted using residual LSTM without Attention block. In addition, we conducted a series of tests employing GRU (Cho et al., 2014) layers in place of LSTM.

The NTU-RGB+D dataset is much larger than the UTD-MHAD and contains data from multiple viewpoints collected using three depth sensors. It is evident from the Tables 6.14, 6.7, and 6.8 that the fusion of features results in a significant increase in the $Accuracy$ of the multi-modal stream (F_Score) for all combinations in both $C-S$ and $C-V$ protocols. In addition, the various score fusion operations result in improved performance compared to a single Softmax score. It is observed that the

Table 6.6: HAR $Accuracy$ (in %) of proposed DL model and results of score fusion operations on UTD-MHAD dataset using $C - S$ protocol. [TA indicated Temporal Attention, Best results are in bold]

Model	S_Score	D_Score	C_Score	$Fusion\_1$	$Fusion\_2$	$Fusion\_3$	$Fusion\_4$	$Fusion\_5$	$Fusion\_6$	$Fusion\_7$	$Fusion\_8$
3 layer LSTM +TA	86.51	78.37	87.9	88.6	88.37	88.13	88.6	89.06	88.6	88.37	**89.76**
2 layer LSTM +TA	83.02	75.11	83.95	85.81	86.51	84.88	88.83	88.83	87.67	88.37	87.67
1 layer LSTM +TA	83.02	77.9	81.62	88.13	85.34	81.39	83.02	84.88	84.18	87.9	87.44
3 layer LSTM	86.5	64.65	85.3	87.9	85.34	85.81	86.74	84.88	84.18	87.9	87.9
2 layer LSTM	82.32	77.2	81.39	86.74	83.95	83.95	84.18	85.81	84.88	88.13	87.44
1 layer LSTM	85.11	74.65	84.18	87.9	87.6	85.34	85.34	85.34	84.88	87.67	87.44
3 layer GRU +TA	84.18	80	86.04	87.67	85.58	86.04	86.04	86.74	85.81	88.6	87.67
2 layer GRU +TA	82.55	77.67	84.18	88.6	85.34	85.81	85.81	86.74	86.04	88.13	88.6
1 layer GRU +TA	80	68.13	81.86	83.48	82.09	81.39	81.16	84.18	82.55	84.88	83.95
3 layer GRU	85.81	74.65	85.34	89	87.9	87.67	87.67	87.67	85.34	89.5	88.3
2 layer GRU	84.88	73.9	85.34	86.27	86.04	85.11	85.81	87.2	86.04	87.67	87.9
1 layer GRU	86.04	73.9	86.97	88.37	86.74	88.37	88.13	86.97	87.44	88.83	88.6

$Fusion\_7$ yields the best performance for both $C - S$ and $C - V$ evaluation on NTU RGB+D. Where as, for UTD-MHAD the $Fusion\_8$ showed optimal performance. The set of experiments with varying numbers of LSTM layers concluded that three layers of residual LSTM with Attention blocks achieved the optimal performance on both single-view and multi-view datasets on different evaluation protocols. Tables 6.14, 6.7, and 6.8 also demonstrates that the use of Attention in LSTM has a very important role in recognizing the actions.

Figures 6.14, and 6.15 depict the confusion matrix obtained for $Fusion\_7$ on 3-layer residual LSTM with Attention on NTU RGB+D dataset using $C - S$ and $C - V$ protocol, respectively. From the figures, it is clear that around 15 and 17 actions are recognized with 95% and more $Accuracy$, respectively, in $C - S$ and $C - V$ proto-

Table 6.7: HAR $Accuracy$ (in $\%$) of proposed DL model and results of score fusion operations on NTU RGB+D dataset using $C - S$ protocol. [TA indicates Temporal Attention, Best results are in bold]

Model	S_Score	D_Score	C_Score	$Fusion\_1$	$Fusion\_2$	$Fusion\_3$	$Fusion\_4$	$Fusion\_5$	$Fusion\_6$	$Fusion\_7$	$Fusion\_8$
3 layer LSTM +TA	69.3	73.55	79.51	82.69	80.48	80.15	79.87	81.83	81.33	**83.8**	82.82
2 layer LSTM +TA	67.57	73.08	78.96	82.21	79.79	79.68	79.22	81.72	81.15	83.15	82.21
1 layer LSTM +TA	65.6	72.96	78.82	81.31	79.16	79.05	79.17	81.38	80.68	82.43	81.73
3 layer LSTM	68.3	73.11	79.26	82.58	80.13	79.49	79.25	82.31	81.56	83.42	82.49
2 layer LSTM	67.45	72.77	79.05	82.03	79.92	79.49	79.23	81.44	81.06	82.96	81.98
1 layer LSTM	65.96	72.95	79.25	81.72	79.22	79.18	79.06	81.58	80.89	82.9	82.02
3 layer GRU +TA	68.28	73.57	79.35	82.47	80.36	79.78	79.37	82.36	81.56	83.72	82.84
2 layer GRU +TA	67.56	73.25	79.4	82.24	80.07	79.73	79.51	81.67	81.11	83.41	82.53
1 layer GRU +TA	65.39	73.64	78.64	81.47	79.42	78.97	78.85	81.24	80.79	82.65	81.78
3 layer GRU	68.39	73.64	78.64	81.84	79.42	78.97	78.85	81.24	80.79	82.65	81.78
2 layer GRU	67.73	73.2	79.18	82.02	80.17	79.67	79.44	81.41	81.16	83.01	82.41
1 layer GRU	66	73.11	79.31	81.84	79.74	79.63	79.39	81.57	80.76	82.66	82.1

cols. Around half of the actions are recognized with minimum of 90% $Accuracy$. The proposed model performed poorly in recognizing only a few actions, such as reading, writing, check time, etc., and greatly affected the overall performance of the proposed work.

Figure 6.16 shows the confusion matrix for the UTD-MHAD dataset. In this, 19 actions were recognized with more than 93% $Accuracy$ (16 with 100% $Accuracy$). The 'wave' (A_3) action is recognized poorly. Most times it is very much confused with 'draw circle_CCW' (A_10) and 'knock' (A_19) actions. So, the proposed model achieved excellent performance in recognizing the majority of the actions, but the poor performance in recognizing one action has a significant effect on the overall performance of the proposed approach.

Table 6.8: HAR $Accuracy$ (in %) of proposed DL model and results of score fusion operations on NTU RGB+D dataset using $C-V$ protocol. [TA indicates Temporal Attention, Best results are in bold]

Model	S_Score	D_Score	C_Score	$Fusion\_1$	$Fusion\_2$	$Fusion\_3$	$Fusion\_4$	$Fusion\_5$	$Fusion\_6$	$Fusion\_7$	$Fusion\_8$
3 layer LSTM +TA	74.62	77.14	86.81	88.06	85.98	87.2	87.24	88.19	87.88	**89.75**	89.15
2 layer LSTM +TA	73.01	77.09	86.8	87.46	85.28	87.19	87.01	87.88	87.55	89.08	88.64
1 layer LSTM +TA	70.47	76.98	85.53	86.59	84.13	85.95	85.87	86.7	86.22	88.34	87.82
3 layer LSTM	73.23	76.89	87.69	87.67	85.38	87.48	87.39	88.27	88.06	89.6	89.28
2 layer LSTM	73.35	76.41	86.75	86.99	84.77	87.13	86.93	87.84	87.42	89.26	88.74
1 layer LSTM	70.44	76.96	86.8	86.51	84.21	86.9	86.79	87.32	87.16	88.7	88.31
3 layer GRU +TA	72.49	76.9	86.6	87.13	84.97	86.84	86.74	87.74	87.48	89.09	88.52
2 layer GRU +TA	72.32	77.05	86.34	86.84	84.68	86.65	86.47	86.87	86.64	88.79	88.3
1 layer GRU +TA	69.68	76.9	86.22	86.16	83.65	86.41	86.42	86.95	86.8	88.66	88.07
3 layer GRU	74.13	76.44	87.28	87.24	84.93	87.26	87.26	88.17	87.88	89.42	88.82
2 layer GRU	73.28	76.9	86.66	87.01	84.84	86.86	86.86	87.75	87.44	89.21	88.6
1 layer GRU	70.6	76.98	86.62	86.48	84.1	86.62	86.59	87.06	87.06	88.54	88.31

We also performed analysis on statistics of individual actions in terms of $F-Score$. Tables 6.9 and 6.10 report the top and bottom five scores in $F-Score$ along with labels for both $C-S$ and $C-V$ protocol on NTU RGB+D and UTD-MHAD datasets, respectively. IN UTD-MHAD dataset 12 actions were recognized with $F-Score$ of 1. The lowest $F-Score$ 0.32 obtained by the 'wave' action. Where as, in NTU RGB+D the lowest $F-Score$ is 0.55 and 0.61 in $C-S$ and $C-V$ protocol respectively. In $C-V$ protocol 17 actions achieved more than 0.96 $F-Score$, and in $C-S$ protocol 11 actions reported above 0.95 $F-Score$. This shows that the proposed system is a promising HAR system for various applications depends on HAR.

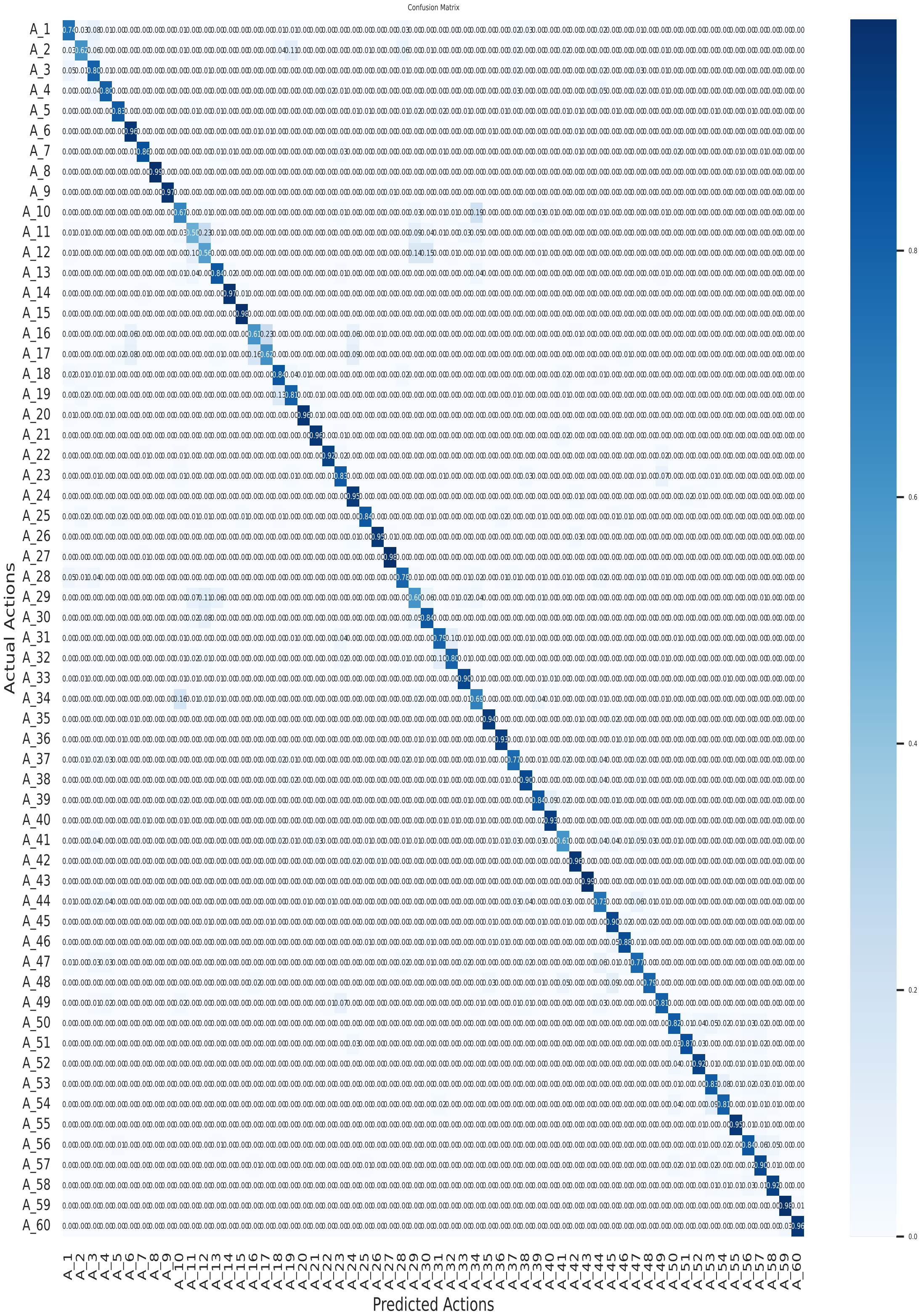


Figure 6.14: Confusion matrix obtained for $C - S$ protocol on NTU RGB+D

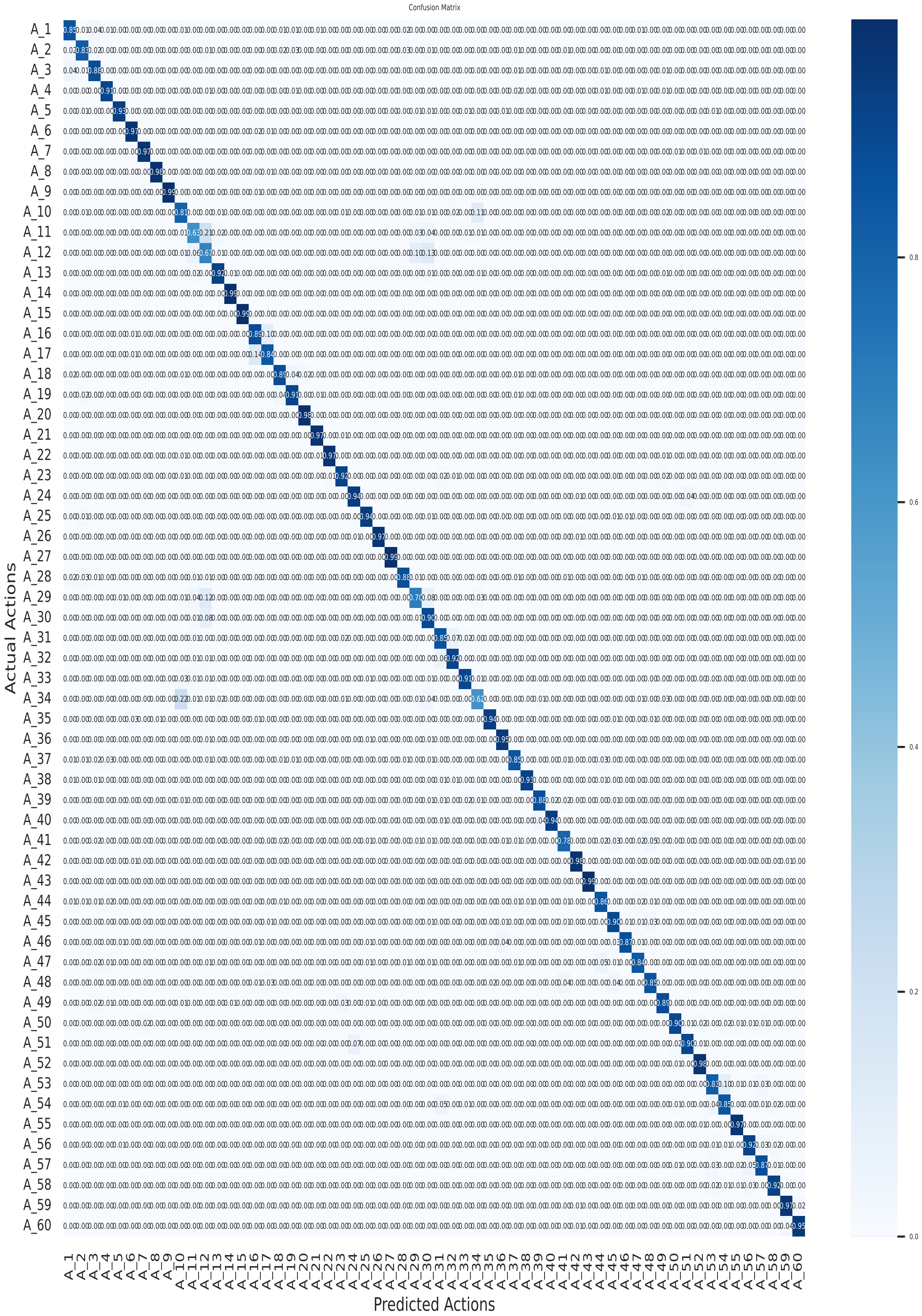
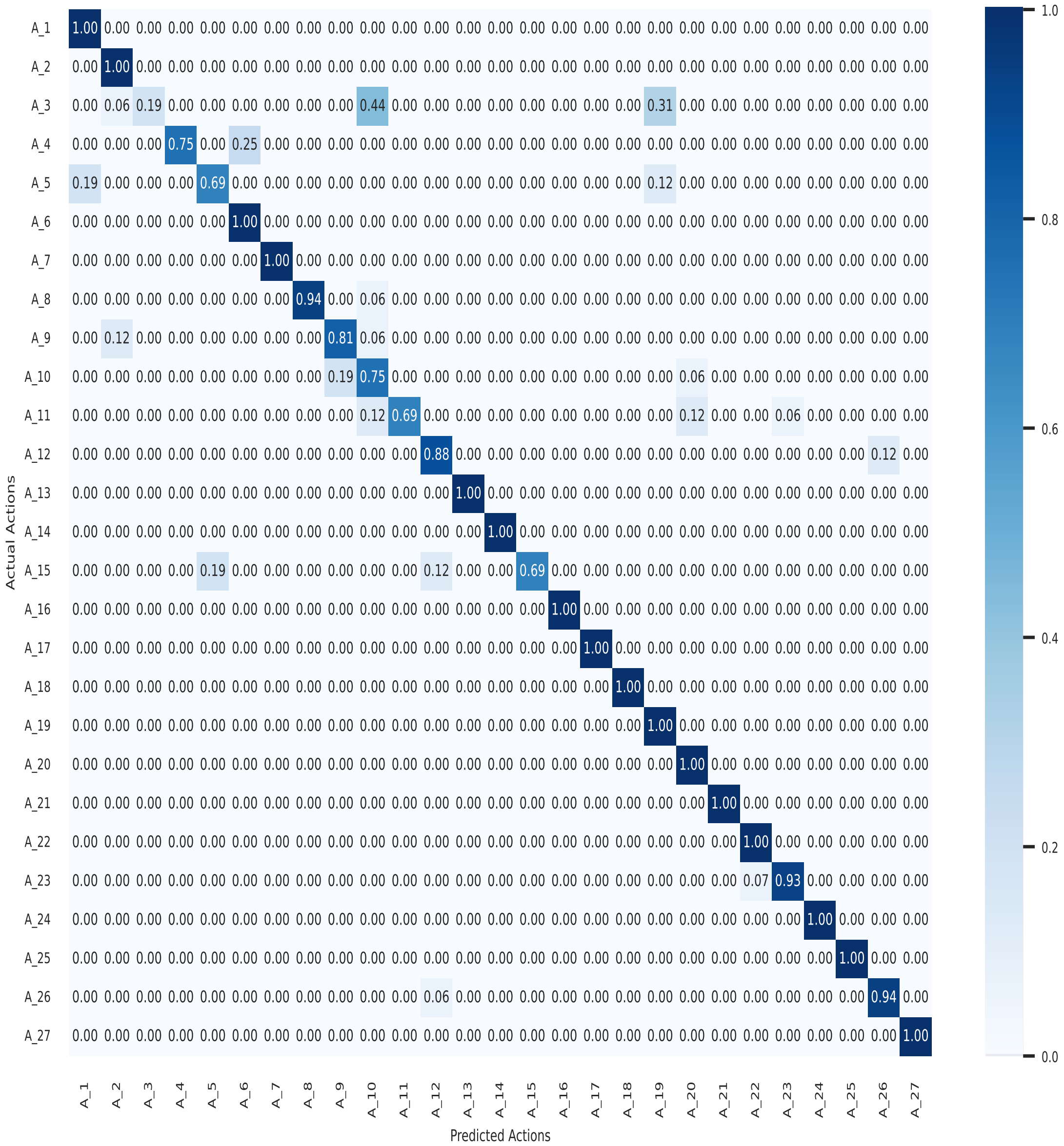


Figure 6.15: Confusion matrix obtained for $C - V$ protocol on NTU RGB+D.



Actual Actions
Predicted Actions
1.0
0.8
0.6
0.4
0.2
0.0

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_10	A_11	A_12	A_13	A_14	A_15	A_16	A_17	A_18	A_19	A_20	A_21	A_22	A_23	A_24	A_25	A_26	A_27
A_1	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A_2	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A_3	0.00	0.06	0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.44	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A_4	0.00	0.00	0.00	0.75	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A_5	0.19	0.00	0.00	0.00	0.69	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A_6	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A_7	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A_8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A_9	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.81	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A_10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.75	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A_11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.69	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.06	0.00	0.00	0.00	0.00
A_12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.88	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.00
A_13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A_14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A_15	0.00	0.00	0.00	0.00	0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.69	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A_16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A_17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A_18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A_19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A_20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
A_21	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
A_22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
A_23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.93	0.00	0.00	0.00	0.00
A_24	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
A_25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
A_26	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.00
A_27	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Figure 6.16: Confusion matrix obtained for $C - S$ protocol on UTD-MHAD.

Ablation Experiments

Several experiments were conducted by removing the LSTM layers from the proposed 3-layer system. Also, experimented with the models without temporal Attention for residual LSTM. Further, another variant of RNN called GRU layers with and without temporal Attention are tried on recognizing the actions. The lower part of the Tables 6.14, 6.7, and 6.8 reports the results of same on UTD-MHAD, and NTU RGB+D using $C - S$ and $C - V$ protocols. In addition, tested the significance of Spatial Attention in

Table 6.9: Top and bottom five $F - Score$ (in %) and corresponding actions recognized from NTU-RGB+D.

	$C - S$ **Protocol**		$C - V$ **Protocol**	
	F-Score	**Action Labels**	**F-Score**	**Action Labels**
Top 5	99	sit down, stand up, falling down	100	take off jacket, falling down
	98	take off jacket, jump up	99	sit down, stand up, put on jacket, jump up
	97	walking towards, waking apart	98	cheer up, hopping
	96	put on jacket, put on hat, hopping	97	throw, put on hat/cap, take off a hat/cap, staggering, pushing, hugging
	95	hugging	96	pick up, walking towards, walking apart
Bottom 5	70	eat meal, headache	75	clapping
	67	put on a shoe, sneeze/cough	73	play with phone/tablet
	66	clapping, take off a shoe, check time (from watch)	69	reading
	61	play with phone/tablet	68	check time (from watch)
	55	reading, writing	61	Writing

depth descriptor of the action. Table 6.11 reports D_Score and C_Score (Classification scores from stream S2 and S3) obtained on all the evaluation protocols on UTD-MHAD and NTU RGB+D datasets without Spatial Attention, and it is proved that it has a significant role in improving the performance of the proposed approach.

6.2.2.4 Comparison with State-of-the-art Works

Finally, the performance of the proposed work is compared with various published works. Table 6.12 reports the comparison of works on UTD-MHAD dataset. Based on skeleton and depth data, the proposed work outperformed the multi-modal work (Kamel et al., 2019) with same combination of data modalities.

Table 6.13 compares the $Accuracy$ of the proposed system with various uni-modal and multi-modal works on NTU RGB+D. The proposed work showed almost equal and better performance than the existing works except (Wu et al., 2022). However, the

Table 6.10: Top and bottom five $F - Score$ (in %) and corresponding actions recognized from UTD-MHAD.

	F-Score	**Action Labels**
Top 5	100	basketball shoot, Boxing, Baseball swing, Arm curl, Tennis serve, Push, Pickup and throw, sit to stand, stand to sit, squat
	97	Draw X, Jog
	93	Walk
	91	Swipe left, Swipe right, Catch
	89	Arm cross
Bottom 5	82	Push
	81	draw triangle, tennis swing, draw circle CW
	73	throw
	62	draw cirlce_CCW
	32	wave

Table 6.11: $Accuracy$ (in %) of depth and fusion stream without Spatial Attention

Models	D_Score	C_Score
UTD-MHAD without SA	42.79	57.44
UTD-MHAD with SA	**76.51**	**85.34**
NTU RGB+D ($C - S$ Protocol) without SA	70.05	78.26
NTU RGB+D ($C - S$ Protocol) with SA	**73.55**	**79.51**
NTU RGB+D ($C - V$ Protocol) without SA	72.01	85.10
NTU RGB+D ($C - V$ Protocol) with SA	**77.14**	**86.81**

data to be processed by these this approach is significantly higher than ours. Other two multi-modal approaches, namely: (El-Ghaish et al., 2018) and (Fan et al., 2020) using a skeleton and RGB data, showed almost equal performance to the proposed work. (El-Ghaish et al., 2018) showed 0.5%, and 0.4% greater $Accuracy$ in $C - S$ and $C - V$ protocol, respectively, than ours. (Fan et al., 2020) achieved slightly less (by .4%) in $C - V$ and high (by .4%) $Accuracy$ than ours. However, the use of RGB data may incur more demand on resources and may get affected by the color of the context.

Table 6.12: Comparison of proposed work's recognition $Accuracy$ (in %) on UTD-MHAD with existing approaches

Method	Modality	Cross-Subject
3D histogram (Zhang et al., 2017)	Depth	84.4
JDM-CNN (Li et al., 2017)	Skeleton	88.1
JTM-CNN (Wang et al., 2018)	Skeleton	87.9
D-CNN (Kamel et al., 2019)	Skeleton+Depth	88.14
MLSL (Yang et al., 2020)	Depth	88.37
EAC (Bulbul and Ali, 2021)	Depth	88.37
SJ-ATP (Sima et al., 2022)	Skeleton	86.37
SML-3DCNN (Wu et al., 2022)	RGB+Depth	93.57
Proposed Work	Skeleton+Depth	89.76

6.3 HI System using Fusion of Gait Data in Skeleton and Depth Modality

The fusion of data modalities to improve the performance of HAR systems by compensating the shortcomings of individual data formats inspired us to propose a work for HI systems by fusing gait data in skeleton and depth formats and DL model. The key contributions of this thesis work are as follows.

Contributions:

- A RGB image representation of the human gait cycle in sequence of skeleton data.
- A image representation of sequence of depth images of a gait cycle in depth data.
- A multi-stream CNN-based DL model to learn features from new representation of gait cycle.

6.3.1 Proposed Methodology

The overall architecture of proposed method for HI system based on the fusion of gait data in skeleton and depth modality is shown in Figure 6.17. The process begins with pre-processing, in which the gait cycle in two different data modalities can be represented as image. Further, the image representations are trained using a multi-stream CNN-based DL model. Finally, the scores from different streams are combined for the final recognition.

Table 6.13: Comparison of proposed work's recognition $Accuracy$ (in %) on NTU RGB+D with existing approaches

Method	**Cross-Subject**	**Cross-View**	**Modality**
JDM-CNN (Li et al., 2017)	76.2	82.3	Skeleton
JTM-CNN (Wang et al., 2018)	76.32	81.08	Skeleton
I-PPM (El-Ghaish et al., 2018)	84.3	90.4	Skeleton+RGB
TSSI+GLAN+SSAN (Yang et al., 2019)	82.4	89.1	Skeleton
MV-Dynamic Images (Xiao et al., 2019)	84.6	87.3	Depth
CACA (Fan et al., 2020)	84.2	89.3	Skeleton+RGB
DS-LSTM (Jiang et al., 2020)	77.80	87.33	Skeleton
TS-MSTD (Dhiman and Vishwakarma, 2020)	79.4	84.1	RGB+Depth
PoT2I + Inception-v3 (Huynh-The et al., 2020)	83.85	90.33	Skeleton
SSI (Shao et al., 2021)	81.9	88.7	Skeleton
SML-3DCNN (Wu et al., 2022)	91.13	94.31	RGB+Depth
M-Att (Li et al., 2022)	83.72	93.80	Skeleton
Proposed Work	83.8	89.75	Skeleton+Depth

6.3.1.1 Pre-processing

The sequence of skeleton and depth frames, denoted by $S = \{s_1, s_2, s_3, ..., s_N\}$ and $D = \{d_1, d_2, d_3, ..., d_N\}$, where, N is the number of frames, are pre-processed using separate steps to convert them into image formats. A detailed explanation of each of these transformation processes is given below.

Skeleton Frame Sequence

Each skeleton frame has a set of 3D coordinates of human body joints. The sequence of skeleton frames is converted to RGB image considering the joint coordinate position in each frame. For this, initially, all the joints from the set of frames are transformed into a global coordinate system making the hip center as the center of the coordinate system is already explained in Section 6.2.1.1 (Page No. 141). Further inspired by the

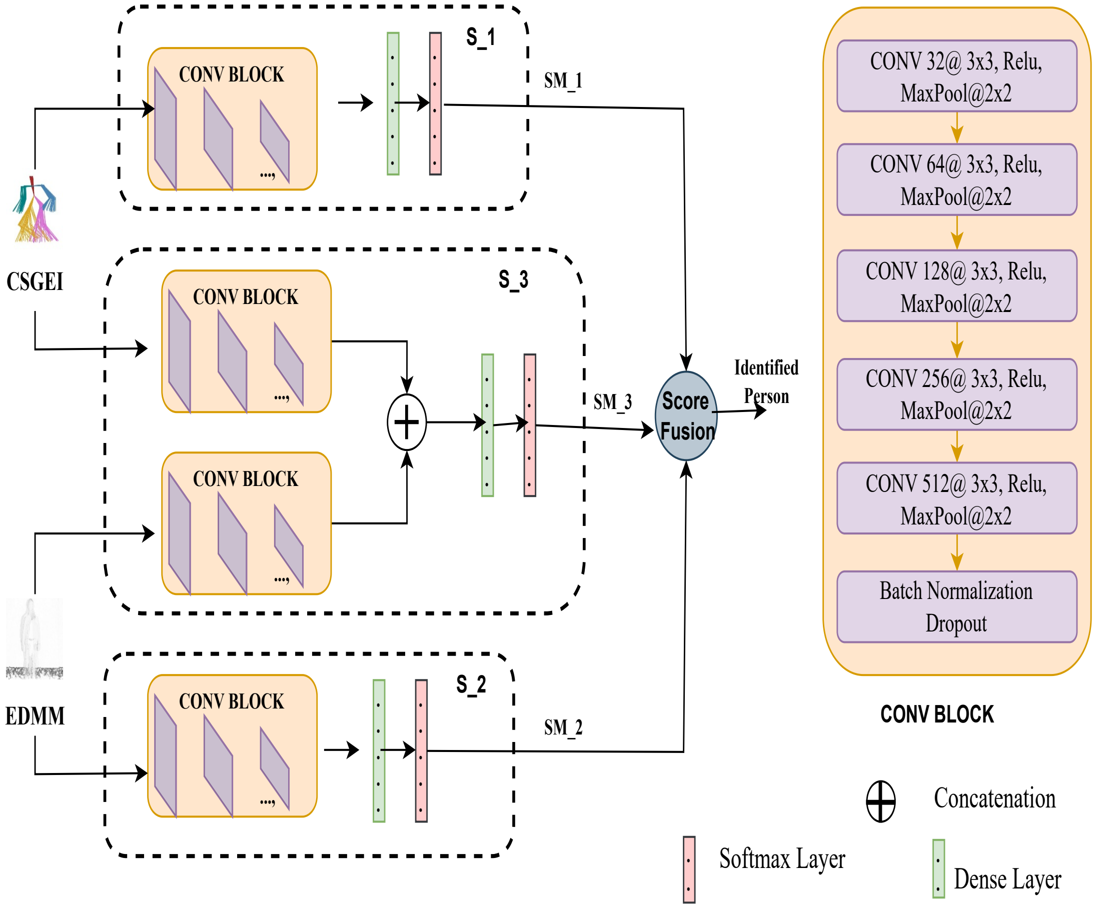
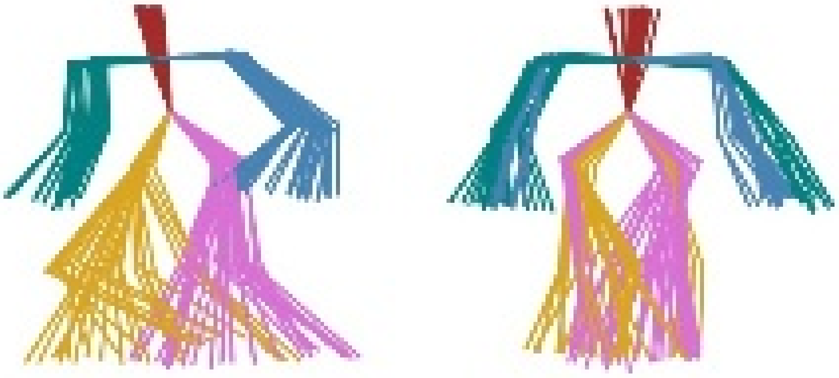

S_1
CONV BLOCK
....,
SM_1
CSGEI
S_3
CONV BLOCK
....,
CONV BLOCK
....,
SM_3
Score Fusion
Identified Person
EDMM
S_2
CONV BLOCK
....,
SM_2
CONV 32@ 3x3, Relu, MaxPool@2x2
CONV 64@ 3x3, Relu, MaxPool@2x2
CONV 128@ 3x3, Relu, MaxPool@2x2
CONV 256@ 3x3, Relu, MaxPool@2x2
CONV 512@ 3x3, Relu, MaxPool@2x2
Batch Normalization Dropout
CONV BLOCK
Concatenation
Softmax Layer
Dense Layer

Figure 6.17: Proposed multi-stream CNN-based deep learning model for HI.

GEI representation of a sequence of silhouette images of the gait cycle, we proposed Color-coded Skeleton Gait Energy Image (CSGEI), where all the skeleton images are combined into a single image frame. So, each skeleton is divided into five body parts, namely: left leg, right leg, left hand, right hand, and trunk. Each of these body parts is color coded with five different colors. The sample CSGEI images generated are shown in Figure 6.18.

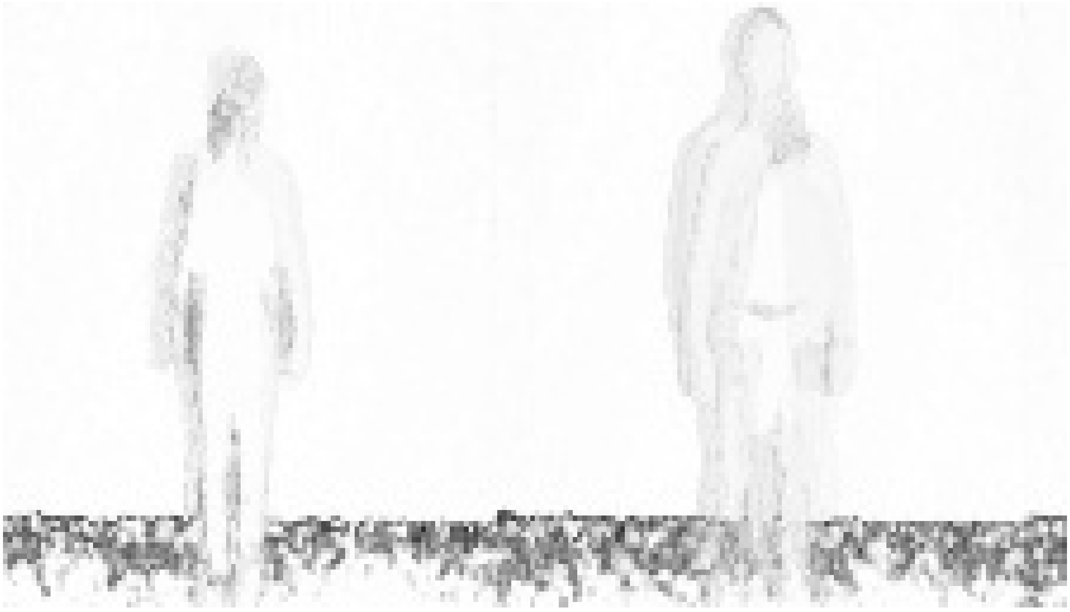
Depth Frame Sequence

The sequence of depth frames in a gait cycle can be transformed to a single image to represent its spatio-temporal features. For this, the EDMM images are generated from the depth frame sequence of a gait cycle using method described in Section 6.2.1.1 (Page No. 141). The sample EDMM images are shown in Figure 6.19.



(a) (b)

Figure 6.18: Sample CSGEI images generated for gait cycles from IAS_Lab gait dataset: (a) Person-1. (b) Person-2.



(a) (b)

Figure 6.19: Sample EDMM images generated for gait cycles from IAS_Lab gait dataset: (a) Person-1. (b) Person-2.

6.3.1.2 Deep Learning Model

Figure 6.17 illustrates the proposed multi-stream CNN-based DL model. Primarily three DL streams are built with CONV blocks. The CSGEI, and EDMM are processed by S_1, and S_3, respectively. The second stream S_2 has two sub-streams which are replica of S_1, and S_3 to process both CSGEI, and EDMM images. The features from two sub-streams are concatenated and further processed by Dense layer. The output of

Dense layers are normalized using BN layer. Further the Softmax layer outputs SM_1, SM_2, and SM_3 of S_1, S_2, and S_3, respectively and these are combined with various score fusion operations.

6.3.2 Experiments, Results, and Analysis

6.3.2.1 Dataset

The experiments are conducted on multi-modal gait dataset IAS-Lab. The features of IAS-Lab are already explained in Chapter 5, Section 5.2.2.1. In this we have used both skeleton and depth data.

6.3.2.2 Experiments, and Results

Table 6.14: $Accuracy$ (in %) of proposed multi-modal approach for HI and results of score fusion operations on IAS_Lab dataset. (Best results are in bold)

Data	S_Score	D_Score	C_Score	$Fusion\_1$	$Fusion\_2$	$Fusion\_3$	$Fusion\_4$	$Fusion\_5$	$Fusion\_6$	$Fusion\_7$	$Fusion\_8$
Testing A	**37**	14	22	24	27	24	27	10	16	33	31
Testing B	**17**	15	7	11	13	7	4	7	4	4	4

The DL model is trained for 500 epochs with initial learning rate of 0.0001. We have used various score fusion operations mentioned in Table 6.5 to combine Softmax scores from different streams of DL model. Table 6.14 shows the results obtained on 'TestingA' and 'TestingB' part of IAS-Lab dataset on various score fusion operations. The results obtained are not encouraging. In both 'TestingA' and 'TestingB' the skeleton stream S_1 performed well as compared to S_2 and S_3. The main reason for this is number of samples in the dataset. As we got few gait cycles (38 after augmentation by flipping got 78 samples for training part), the DL model failed to learn the features.

6.4 Limitations

- Experimental results demonstrated that the proposed system performed exceptionally well in recognizing most actions. However, it demonstrated poor performance when recognizing only a few actions. Most of the time, the analysis of poorly performing actions revealed that they share a similar structure when summarizing the data sequence. The variance in number of sub-actions in actions are not considered. Adding multi-modal data not only improved the performance, but also increased the computations required to manage the different data types in a multi-stream DL model significantly.

- The results on HI system are not encouraging. The main reason is the CNN-based DL model is trained with few features. Also, the image representation of the gait cycle may not perform well in representing spatio-temporal features.

6.5 Summary

Two different approaches are proposed for HAR using a combination of skeleton and depth data of human action. In the first work, efficient action descriptors using skeleton and depth data are formed and trained using the multi-stream DL model. Performance is measured using two widely-used datasets for HAR. Additionally, several experimental setups established the importance of the Attention unit in accurate recognition. In the second approach, we attempted to investigate the sub-actions comprising human actions. To reduce the number of features to be processed with less number of computations, we proposed the novel depth descriptor MS-EDMM, in which the spatio-temporal features involved in sub-actions of action in depth sequences are summed and represented as a single image. To process the skeleton data sequence, we categorized the skeleton joints into five groups, aggregated them up, and trained with the proposed multi-stream DL model. Using standard evaluation protocols, the proposed approach is evaluated using the benchmark single-view and multi-view datasets. The experimental results proved the efficiency of the proposed approach using the fusion of depth and skeleton data for HAR.

The success in improving the performance of HAR system encouraged the use of multi-modal gait data to propose the HI system. Further, this chapter proposes new image representations of the gait cycles from skeleton and depth data. Also, the images are trained with CNN-based multi-stream DL model. The experimental results on one publicly available multi-modal gait dataset are not promising. As the dataset is small, the proposed model failed to learn the features from images and hence we did not achieve encouraging performance.

Chapter 7

Conclusions and Future Directions

7.1 Conclusions

In recent years, the proliferation of smart environments, such as smart cities, smart campuses, smart transportation, and smart parking, aims to provide people with a secure, healthy, enjoyable, and comfortable life. This results in deploying and integrating various sensor types to automate various tasks. Ongoing research is being conducted in various areas of smart environments for a variety of applications, including smart waste management, smart energy management, etc. Smart surveillance is one of the most critical applications in smart environments, with a wide range of applications. The tasks of a smart surveillance system include abandoned object detection, alert systems based on events in the scene, monitoring of the activities of people, and so on.

To provide a secure environment, automating monitoring people's activities entails primarily two tasks: recognizing people's actions and identifying people of interest in the scene from a distance. Consequently, this thesis aims to strengthen the unobtrusive HAR and HI systems. This requires content interpretation based on the analysis of images/video captured by cameras. Due to advancements in camera technology, visual data is now accessible in various modalities. Consequently, this work explored the pros and cons of various visual data modalities for HAR and HI tasks. In addition, the shortfalls of HAR and HI from various perspectives are also identified and discussed.

Vision-based HAR is one of the most alluring research areas in the current research community. In addition to smart surveillance, it has a variety of other applications, including sports analysis, entertainment, etc. This thesis examined the existing literature on on HAR and HI through various data modalities and strategies. This thesis makes contributions to the field of HAR and HI based on its review of the existing literature. We conclude the salient features of all thesis contributions as follows.

- The ***first contribution*** of this thesis is on creating context-specific RGB data-based HAR. Human actions vary from domain to domain. With this in mind, a dataset of human actions is compiled to identify students' actions within the computer laboratories of smart campuses. Due to the need for unobtrusive monitoring of individuals, the dataset is comprised of image frames from spontaneous video captures by CCTV cameras installed in various locations within computer laboratories. We presented the DL model based on transfer learning and YOLOv3 for recognizing student actions in computer laboratories. The model can identify and localize multiple actions from RGB images. In addition, this work introduced a technique for reducing the number of frames in a video based on selecting key frames using a proposed template matching scheme. This significantly reduces the number of frames that must be analyzed and stored for future use. This allowed only the essential video content to be stored rather than the entire video. The key limitations of this thesis work are as follows:

  - It can recognize only simple actions.
  - The method uses only spatial features, so, the actions with temporal features cannot be recognized.
  - The number of annotators are less.

- The ***second contribution*** of this thesis is the skeleton data-based HAR system. In this effective tree-like representation of the skeleton joints making, the hip center as the root and DFS traversal of nodes is introduced. The distance features are computed from this new representation and processed using a DL model with Dense layers. Finally, the identified actions are determined using a majority voting scheme based on Softmax scores for every frame. The experimental evaluation of skeleton-data-based HAR on a single-view dataset revealed its promising performance. The key limitations of this work is as follows:
  - The actions are recognized by frame-wise classifications, but temporal features between the frames are not considered.
  - The noisy frames will greatly effect the performance as initial classification is based on frame level information.

- The ***third contribution*** of this thesis is unobtrusive human identification. Two works on novel gait event-specific features and advanced DL models are proposed. At first, the optimal sets of features are extracted based on inter/intra frame skeleton joint distance and angles. The quantitative summary of features is computed with an emphasis on reducing the effect of noise and occlusion on recognition performance. This reduces the number of features that must be processed compared to approaches that employ skeleton-wise features. As gait is a cyclical movement of human limbs with a unique pattern for each individual, temporal characteristics within the sequence of gait events play a crucial role in classification. Therefore, an LSTM/GRU-based DL model is proposed for learning the quantitative summary of features extracted from gait events. The proposed DL model has significantly fewer total parameters than existing methods. In addition, evaluating the proposed system on benchmark single/multi-view datasets using a variety of state-of-the-art evaluation protocols demonstrated its superior performance in HI. The CMC test demonstrated that the proposed approaches reached more than 95% $Accuracy$ in lower level ranks on all the datasets. The key limitations of this thesis work are as follows:

  - Though there is advantage in considering the quantitative summary of features, the benefits of features in correct frames are suppressed.
  - The multi-view dataset is a small scale dataset. So, it must be tested in large-scale multi-view dataset.

- The ***fourth contribution*** of this thesis is single/multi-view HAR and HI system using multi-modal data. In this, two works are proposed using 3D skeleton and depth data from Kinect depth sensors for HAR. Both works introduced novel action representations and multi-stream DL models. In addition, a small work is proposed for HI using fusion of gait data in skeleton and depth formats. The first work on HAR maps the 3D coordinates and orientation features from a sequence of skeleton frames to single RGB image. Additionally, the motion in a series of depth frames is condensed into a single image. Further, the multi-stream DL model with CNN and LSTM with Attention units process the action representations in image format. The test on two single-view human action datasets proved its performance. The second work on HAR effectively explores the benefit of using overlapped sub-actions. The sequence of depth frames is divided into four sub-actions, and the depth maps of overlapping sub-actions are condensed into a single image that represents the action. Similarly, the skeleton sequence is grouped into ten overlapped sub-actions. Further, each skeleton is divided into five regions. The quantitative summary of features from each region is extracted for each sub-action. The extracted features are trained using multi-stream DL model with CNN and spatial Attention, and LSTM/GRU with temporal Attention units. Proposed a set of score fusion operations that use the scores from DL streams processing different data modalities to improve performance. Results demonstrated its superior performance on single and multi-view datasets.

In the proposed HI system based on the fusion of depth and skeleton data, we initially represented the sequence of skeletons as a RGB image. Which combines all the skeletons into one image. Also, the sequence of depth images of the gait cycle is summarized into a single image. The two types of images are trained with a multi-stream CNN-based DL model. The experiment on a small-scale multi-modal gait dataset demonstrated poor performance. The key limitations of this thesis work are as follows:

- The number of sub-actions in different actions varies, but same number of sub-actions are considered for all actions.
- The overall performance of the system is greatly affected by few actions which has similar movement when observed from far.
- The number of samples used to train the multi-stream DL model is much less in the proposed multi-modal HI system, leading to poor performance.
- The proposed image representations for HAR/HI system must be improved to capture spatio-temporal features of the gait cycle.

7.2 Future Directions

This research work carried out in thesis has a number of future directions that can be pursued to enhance the work in both HAR and gait-based HI domains to support smart applications. The details are as follows.

Human Action Recognition Systems

- As human actions vary based on context, developing domain-specific datasets and RGB-based HAR systems is necessary.
- Generate a domain-specific dataset of abnormal actions and HAR systems so that it can be used in alert systems to alert individuals/authorities about abnormal activities.
- Create a domain-specific dataset of human action videos and develop HAR systems capable of learning temporal aspects of various actions.

- Develop an HAR system capable of recognizing and localizing/segmenting the various actions in large videos.

- Utilize the context information of the scene to improve the performance of the RGB-based HAR system by identifying the objects in the scene for action recognition at a finer level of detail.

- The skeleton data-based action recognition can be enhanced by defining an action representation that takes temporal aspects of the skeleton sequence into account. There is a need of defining classification models that can learn the dynamics present in an action's representation, and also developing DL models that focus on the most important features of human actions. Additionally, these models are to be tested using multi-view action data.

- The multi-modal HAR system can be improved by defining action descriptors that can capture spatio-temporal features that can differentiate between extremely similar actions.

- The number of sub-actions varies from action to action. Therefore, rather than a fixed number of sub-actions, methods should be developed for automatically locating the sub-action regions in an action video.

- If contextual features available in RGB data are integrated for feature extraction, then the performance may be enhanced despite increased computing and resource requirements.

- Develop systems to continuously recognize actions and summarize the observed actions for future use by integrating HAR systems with natural language processing in a context where there is an abundance of visual data that is difficult to analyze and summarize.

- In addition, using federated learning approach, implement and test both HAR using different data modalities to make them suitable for the evolving needs of smart-environments. Use the evolutionary DL model for HAR to optimize the computations further. Also, create and test lightweight models for HAR to make it suitable for use in resource constraint scenarios.

- Develop advanced HAR systems to recognize more complex actions like actions involving group of people and events.

Human Identification Systems

- The HI systems with the skeleton-data based gait can be further enhanced in several ways. Generate more effective set of gait-event specific features. Additionally, combine gait-event-specific features with global features that pertain to the entire gait cycle in order to boost the performance.

- Utilize various types of Attentions at both the global and event levels to further enhance performance.

- Create multi-modal 3D large scale dataset, so that multi-modal systems can be developed using advanced DL models for multi-view HI to support current scenario in smart environments.

- Develop new representation of gait cycles which can capture spatio-temporal features of gait cycle from arbitrary views in different modalities.

- Improve the performance of depth data based gait recognition by developing approaches for view independent representations.

- Create and test lightweight models for HI to make it suitable for use in resource constraint scenarios.

- Use federated learning approach, implement and test HI systems for the evolving needs of smart-environments.

- Predict the intention of human based on action analysis

References

Abu-Bakar, S. A. R. (2019). "Advances in human action recognition: An updated survey". *IET Image Processing*, *13*(13), 2381–2394.

Agahian, S., Negin, F., and Köse, C. (2020). "An efficient human action recognition framework with pose-based spatiotemporal features". *Engineering Science and Technology, an International Journal*, *23*(1), 196–203.

Aggarwal, J. K. and Ryoo, M. S. (2011). "Human activity analysis: A review". *ACM Computing Surveys (CSUR)*, *43*(3), 1–43.

Aggarwal, J. K. and Xia, L. (2014). "Human activity recognition from 3D data: A review". *Pattern Recognition Letters*, *48*, 70–80.

Ahmad, T., Jin, L., Zhang, X., Lai, S., Tang, G., and Lin, L. (2021). "Graph convolutional neural network for human action recognition: A comprehensive survey". *IEEE Transactions on Artificial Intelligence*, *2*(2), 128–145.

Ahmad, Z., Illanko, K., Khan, N., and Androutsos, D. (2019). "Human action recognition using convolutional neural network and depth sensor data". In *Proceedings of the International Conference on Information Technology and Computer Communications*, 1–5.

Amor, B. B., Su, J., and Srivastava, A. (2016). "Action recognition using rate-invariant analysis of skeletal shape trajectories". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(1), 1–13.

Andersson, V. and Araujo, R. (2015). "Person identification using anthropometric and gait data from kinect sensor". In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Angelini, F., Fu, Z., Long, Y., Shao, L., and Naqvi, S. M. (2019). "2D pose-based real-time human action recognition with occlusion-handling". *IEEE Transactions on Multimedia*, 1–1.

Annadani, Y., Rakshith, D. L., and Biswas, S. (2016). "Sliding dictionary based sparse representation for action recognition". *CoRR*, *abs/1611.00218*.

Ball, A., Rye, D., Ramos, F., and Velonaki, M. (2012). "Unsupervised clustering of people from skeleton data". In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 225–226.

Bari, A. H. and Gavrilova, M. L. (2022). "Kinectgaitnet: Kinect-based gait recognition using deep convolutional neural network". *Sensors*, *22*(7), 2631.

Bari, A. S. M. H. and Gavrilova, M. L. (2019). "Artificial neural network based gait recognition using kinect sensor". *IEEE Access*, *7*, 162708–162722.

Batchuluun, G., Yoon, H. S., Kang, J. K., and Park, K. R. (2018). "Gait-based human identification by combining shallow convolutional neural network-stacked long short-term memory and deep convolutional neural network". *IEEE Access*, *6*, 63164–63186.

Bian, C., Zhang, Y., Yang, F., Bi, W., and Lu, W. (2019). “Spontaneous facial expression database for academic emotion inference in online learning”. *IET Computer Vision*, *13*(3), 329–337.

Bobick, A. and Davis, J. (2001). “The recognition of human movement using temporal templates”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(3), 257–267.

Bobillo, F., Dranca, L., and Bernad, J. (2017). “A fuzzy ontology-based system for gait recognition using kinect sensor”. In *Proceedings of the International Conference on Scalable Uncertainty Management*, 397–404.

Bosch, N. and D’Mello, S. (2019). “Automatic detection of mind wandering from video in the lab and in the classroom”. *IEEE Transactions on Affective Computing*, 1–1.

Boyd, J. E. and Little, J. J. (2005). “Biometric gait recognition”. In Tistarelli, M., Bigun, J., and Grosso, E. (Eds.), *Advanced Studies in Biometrics: Summer School on Biometrics, Alghero, Italy, June 2-6, 2003. Revised Selected Lectures and Papers*, 19–42. Springer.

Brownlee, J. (2018), “How and when to use ROC curves and precision-recall curves for classification in python”. `https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/`. Accessed: 14-05-2019.

Bulbul, M. F. and Ali, H. (2021). “Gradient local auto-correlation features for depth human action recognition”. *SN Applied Sciences*, *3*(5), 1–13.

Candra Kirana, K., Wibawanto, S., and Wahyu Herwanto, H. (2018). “Facial emotion recognition based on viola-jones algorithm in the learning environment”. In *Proceedings of the International Seminar on Application for Technology of Information and Communication*, 406–410.

Cartucho (2018), “mAP (mean average precision)”. `https://github.com/Cartucho/mAP`. Accessed: 12-06-2020.

Cartucho, J., Ventura, R., and Veloso, M. (2018). “Robust object recognition through symbiotic deep learning in mobile robots”. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2336–2341.

Chai, Y., Ren, J., Wenying Han, and Haifeng Li (2011). “Human gait recognition: Approaches, datasets and challenges”. In *Proceedings of the 4th International Conference on Imaging for Crime Detection and Prevention 2011 (ICDP 2011)*, 1–6.

Chen, C., Jafari, R., and Kehtarnavaz, N. (2015). “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)*, 168–172.

Chen, Y., Xia, S., Zhao, J., Zhou, Y., Niu, Q., Yao, R., Zhu, D., and Chen, H. (2022). "Adversarial learning-based skeleton synthesis with spatial-channel attention for robust gait recognition". *Multimedia Tools and Applications*, 1–16.

Cheng, J., Ren, Z., Zhang, Q., Gao, X., and Hao, F. (2022). "Cross-modality compensation convolutional neural networks for rgb-d action recognition". *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(3), 1498–1509.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). "On the properties of neural machine translation: Encoder-decoder approaches". *arXiv preprint arXiv:1409.1259*.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). "Learning phrase representations using RNN encoder-decoder for statistical machine translation". *arXiv preprint arXiv:1406.1078*.

Choi, S., Kim, J., Kim, W., and Kim, C. (2019). "Skeleton-based gait recognition via robust frame-level matching". *IEEE Transactions on Information Forensics and Security*, *14*(10), 2577–2592.

Chollet, F. (2015), Keras. `https://github.com/fchollet/keras`.

Cohen, J. (1960). "A coefficient of agreement for nominal scales". *Educational and Psychological Measurement*, *20*(1), 37–46.

Davis, J. and Goadrich, M. (2006). "The relationship between precision-recall and roc curves". In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, 233–240.

Delaitre, V., Laptev, I., and Sivic, J. (2010). "Recognizing human actions in still images: a study of bag-of-features and part-based representations". In *Proceedings of the 21st British Machine Vision Conference*.

Deng, M. and Wang, C. (2019). "Human gait recognition based on deterministic learning and data stream of microsoft kinect". *IEEE Transactions on Circuits and Systems for Video Technology*, *29*(12), 3636–3645.

Devanne, M., Wannous, H., Berretti, S., Pala, P., Daoudi, M., and Del Bimbo, A. (2013). "Space-time pose representation for 3D human action recognition". In *Proceedings of the International Conference on Image Analysis and Processing*, 456–464.

Dhiman, C. and Vishwakarma, D. K. (2020). "View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics". *IEEE Transactions on Image Processing*, *29*, 3835–3844.

Ding, W., Ding, C., Li, G., and Liu, K. (2021). "Skeleton-based square grid for human action recognition with 3D convolutional neural network". *IEEE Access*, *9*, 54078–54089.

Du, S., Meng, F., and Gao, B. (2016). “Research on the application system of smart campus in the context of smart city”. In *Proceedings of the 8th International Conference on Information Technology in Medicine and Education (ITME)*, 714–718.

Du, Y., Fu, Y., and Wang, L. (2016). “Representation learning of temporal dynamics for skeleton-based action recognition”. *IEEE Transactions on Image Processing*, *25*(7), 3010–3022.

Du, Y., Wang, W., and Wang, L. (2015). “Hierarchical recurrent neural network for skeleton based action recognition”. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1110–1118.

E, S. K. R. (2020). “Shortwave infrared-based phenology index method for satellite image land cover classification”. *Soft Computing for Problem Solving. Advances in Intelligent Systems and Computing,Springer*, *1057*.

El-Ghaish, H., Hussien, M. E., Shoukry, A., and Onai, R. (2018). “Human action recognition based on integrating body pose, part shape, and motion”. *IEEE Access*, *6*, 49040–49055.

Evangelidis, G., Singh, G., and Horaud, R. (2014). “Skeletal quads: Human action recognition using joint quadruples”. In *Proceedings of the 22nd International Conference on Pattern Recognition*, 4513–4518.

Fan, Y., Weng, S., Zhang, Y., Shi, B., and Zhang, Y. (2020). “Context-aware cross-attention for skeleton-based human action recognition”. *IEEE Access*, *8*, 15280–15290.

Firman, M. (2016). “RGBD datasets: Past, present and future”. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 661–673.

Foroughi, H., Aski, B. S., and Pourreza, H. (2008). “Intelligent video surveillance for monitoring fall detection of elderly in home environments”. In *Proceedings of the 11th International Conference on Computer and Information Technology*, 219–224.

Gnouma, M., Ladjailia, A., Ejbali, R., and Zaied, M. (2019). “Stacked sparse autoencoder and history of binary motion image for human activity recognition”. *Multimedia Tools and Applications*, *78*(2), 2157–2179.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). “Deep learning”. The MIT Press.

Gu, J. (2019), “Bbox-label-tool”. `https://github.com/jxgu1016/BBox-Label-Tool-Multi-Class`. Accessed: 08-02-2019.

Gu, Y., Ye, X., Sheng, W., Ou, Y., and Li, Y. (2020). “Multiple stream deep learning model for human action recognition”. *Image and Vision Computing*, *93*, 103818.

Guo, G. and Lai, A. (2014). “A survey on still image based human action recognition”. *Pattern Recognition*, *47*(10), 3343 – 3361.

Hampapur, A., Brown, L., Connell, J., Pankanti, S., Senior, A., and Tian, Y. (2003). "Smart surveillance: Applications, technologies and implications". In *Proceedings of the 4th International Conference on Information, Communications and Signal Processing, 2003 and the Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint*, volume 2, 1133–1138.

Han, J. and Bhanu, B. (2006). "Individual recognition using gait energy image". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(2), 316–322.

Han, J., Shao, L., Xu, D., and Shotton, J. (2013). "Enhanced computer vision with microsoft kinect sensor: A review". *IEEE Transactions on Cybernetics*, *43*(5), 1318–1334.

Haque, A., Alahi, A., and Fei-Fei, L. (2016). "Recurrent attention models for depth-based person identification". In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1229–1238.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition". In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hoang, V., Hoang, D., and Hieu, C. (2018). "Action recognition based on sequential 2D-CNN for surveillance systems". In *Proceedings of the 44th Annual Conference of the IEEE Industrial Electronics Society*, 3225–3230.

Hochreiter, S. and Schmidhuber, J. (1997). "Long short-term memory". *Neural computation*, *9*(8), 1735–1780.

Hosni, N. and Amor, B. B. (2020). "A geometric convnet on 3D shape manifold for gait recognition". In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 3725–3734.

Hou, Y., Li, Z., Wang, P., and Li, W. (2018). "Skeleton optical spectra-based action recognition using convolutional neural networks". *IEEE Transactions on Circuits and Systems for Video Technology*, *28*(3), 807–811.

Huang, X. and Boulgouris, N. V. (2012). "Gait recognition with shifted energy image and structural feature extraction". *IEEE Transactions on Image Processing*, *21*(4), 2256–2268.

Huynh-The, T., Hua, C.-H., Ngo, T.-T., and Kim, D.-S. (2020). "Image representation of pose-transition feature for 3D skeleton-based action recognition". *Information Sciences*, *513*, 112–126.

Huynh-The, T., Hua, C.-H., Tu, N. A., and Kim, D.-S. (2020). "Learning 3D spatiotemporal gait feature by convolutional network for person identification". *Neurocomputing*, *397*, 192–202.

Ince, O. F., Ince, I. F., Park, J. S., and Song, J. K. (2017). “Gait analysis and identification based on joint information using rgb-depth camera”. In *Proceedings of the 14th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 561–563.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). “An introduction to statistical learning”, volume 112. Springer.

Jan C. van Gemert, Mihir Jain, E. G. and Snoek., C. G. M. (2015). “APT: Action localization proposals from dense trajectories”. In *Proceedings of the British Machine Vision Conference (BMVC)*, 177.1–177.12.

Ji, Y., Zhan, Y., Yang, Y., Xu, X., Shen, F., and Shen, H. T. (2020). “A context knowledge map guided coarse-to-fine action recognition”. *IEEE Transactions on Image Processing*, *29*, 2742–2752.

Jiang, X., Xu, K., and Sun, T. (2020). “Action recognition scheme based on skeleton representation with ds-lstm network”. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(7), 2129–2140.

Ju Han and Bir Bhanu (2006). “Individual recognition using gait energy image”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(2), 316–322.

Justus, D., Brennan, J., Bonner, S., and McGough, A. S. (2018). “Predicting the computational cost of deep learning models”. *CoRR*, *abs/1811.11880*.

Kamel, A., Sheng, B., Yang, P., Li, P., Shen, R., and Feng, D. D. (2019). “Deep convolutional neural networks for human action recognition using depth maps and postures”. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *49*(9), 1806–1819.

Karianakis, N., Liu, Z., Chen, Y., and Soatto, S. (2017). “Person depth reid: Robust person re-identification with commodity depth sensors”. *CoRR*, *abs/1705.09882*.

Kastaniotis, D., Theodorakopoulos, I., Economou, G., and Fotopoulos, S. (2013). “Gait-based gender recognition using pose information for real time applications”. In *Proceedings of the 18th International Conference on Digital Signal Processing (DSP)*, 1–6.

Kastaniotis, D., Theodorakopoulos, I., Economou, G., and Fotopoulos, S. (2016). “Gait based recognition via fusing information from euclidean and riemannian manifolds”. *Pattern Recognition Letters*, *84*, 245–251.

Kastaniotis, D., Theodorakopoulos, I., Theoharatos, C., Economou, G., and Fotopoulos, S. (2015). “A framework for gait-based recognition using kinect”. *Pattern Recognition Letters*, *68*, 327 – 335. Special Issue on Soft Biometrics.

Ke, Q., Bennamoun, M., An, S., Sohel, F., and Boussaid, F. (2017). “A new representation of skeleton sequences for 3D action recognition”. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3288–3297.

Ke, Q., Bennamoun, M., An, S., Sohel, F., and Boussaid, F. (2018). “Learning clip representations for skeleton-based 3D action recognition”. *IEEE Transactions on Image Processing*, *27*(6), 2842–2855.

Khamsemanan, N., Nattee, C., and Jianwattanapaisarn, N. (2018). “Human identification from freestyle walks using posture-based gait feature”. *IEEE Transactions on Information Forensics and Security*, *13*(1), 119–128.

Khurana, R. and Singh Kushwaha, A. K. (2018). “A deep survey on human activity recognition in video surveillance”. In *Proceedings of the International Conference on Research in Intelligent and Computing in Engineering (RICE)*, 1–5.

Kim, T. S. and Reiter, A. (2017). “Interpretable 3D human action analysis with temporal convolutional networks”. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPRW)*, 1623–1631. IEEE.

Kingma, D. P. and Ba, J. (2014). “Adam: A method for stochastic optimization”. *arXiv preprint arXiv:1412.6980.*

Ko, B., Hong, J., and Nam, J.-Y. (2015). “Human action recognition in still images using action poselets and a two-layer classification model”. *Journal of Visual Languages and Computing*, *28*, 163 – 175.

Kumar, M., Singh, N., Kumar, R., Goel, S., and Kumar, K. (2021). “Gait recognition based on vision systems: A systematic survey”. *Journal of Visual Communication and Image Representation*, *75*, 103052.

Landis, J. R. and Koch, G. G. (1977). “The measurement of observer agreement for categorical data”. *Biometrics*, *33*(1), 159–174.

Laptev, I. and Lindeberg, T. (2005). “On space-time interest points”. *International journal of computer vision*, *64*(2-3), 107–124.

Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008). “Learning realistic human actions from movies”. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–8.

Li, B., Li, X., Zhang, Z., and Wu, F. (2019). “Spatio-temporal graph routing for skeleton-based action recognition”. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8561–8568.

Li, C., Cui, Z., Zheng, W., Xu, C., Ji, R., and Yang, J. (2018). “Action-attending graphic neural network”. *IEEE Transactions on Image Processing*, *27*(7), 3657–3670.

Li, C., Hou, Y., Wang, P., and Li, W. (2017). “Joint distance maps based action recognition with convolutional neural networks”. *IEEE Signal Processing Letters*, *24*(5), 624–628.

Li, C., Hou, Y., Wang, P., and Li, W. (2019). “Multiview-based 3D action recognition using deep networks”. *IEEE Transactions on Human-Machine Systems*, *49*(1), 95–104.

Li, C., Xie, C., Zhang, B., Han, J., Zhen, X., and Chen, J. (2022). “Memory attention networks for skeleton-based action recognition”. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(9), 4800–4814.

Li, J., Qi, L., Zhao, A., Chen, X., and Dong, J. (2017). “Dynamic long short-term memory network for skeleton-based gait recognition”. In *Proceedings of the IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, 1–6.

Li, M., Leung, H., and Shum, H. P. (2016). “Human action recognition via skeletal and depth based feature fusion”. In *Proceedings of the 9th International Conference on Motion in Games*, 123–132.

Li, W., Nie, W., and Su, Y. (2018). “Human action recognition based on selected spatio-temporal features via bidirectional LSTM”. *IEEE Access*, *6*, 44211–44220.

Li, W., Zhang, Z., and Liu, Z. (2010). “Action recognition based on a bag of 3D points”. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 9–14.

Li, X., Makihara, Y., Xu, C., Yagi, Y., Yu, S., and Ren, M. (2020). “End-to-end model-based gait recognition”. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*.

Li, X., Maybank, S. J., Yan, S., Tao, D., and Xu, D. (2008). “Gait components and their application to gender recognition”. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *38*(2), 145–155.

Liang, C., Chen, E., Qi, L., and Guan, L. (2016). “Improving action recognition using collaborative representation of local depth map feature”. *IEEE Signal Processing Letters*, *23*(9), 1241–1245.

Liang, C., Liu, D., Qi, L., and Guan, L. (2020). “Multi-modal human action recognition with sub-action exploiting and class-privacy preserved collaborative representation learning”. *IEEE Access*, *8*, 39920–39933.

Liang, C., Qi, L., He, Y., and Guan, L. (2018). “3D human action recognition using a single depth feature and locality-constrained affine subspace coding”. *IEEE Transactions on Circuits and Systems for Video Technology*, *28*(10), 2920–2932.

Liang Wang, Huazhong Ning, Tieniu Tan, and Weiming Hu (2004). “Fusion of static and dynamic body biometrics for gait recognition”. *IEEE Transactions on Circuits and Systems for Video Technology*, *14*(2), 149–158.

Liao, R., Li, Z., Bhattacharyya, S. S., and York, G. (2022). “PoseMapGait: A model-based gait recognition method with pose estimation maps and graph convolutional networks”. *Neurocomputing*, *501*, 514–528.

Liao, R., Yu, S., An, W., and Huang, Y. (2020). “A model-based gait recognition method with body pose and human prior knowledge”. *Pattern Recognition*, *98*, 107069.

Limcharoen, P., Khamsemanan, N., and Nattee, C. (2020). “View-independent gait recognition using joint replacement coordinates (jrcs) and convolutional neural network”. *IEEE Transactions on Information Forensics and Security*, *15*, 3430–3442.

Limcharoen, P., Khamsemanan, N., and Nattee, C. (2021). “Gait recognition and re-identification based on regional LSTM for 2-second walks”. *IEEE Access*, *9*, 112057–112068.

Liu, J., Shahroudy, A., Perez, M. L., Wang, G., Duan, L., and Kot Chichung, A. (2019). “NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1.

Liu, J., Shahroudy, A., Xu, D., Kot, A. C., and Wang, G. (2018). “Skeleton-based action recognition using spatio-temporal LSTM network with trust gates”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(12), 3007–3021.

Liu, J., Wang, G., Duan, L.-Y., Abdiyeva, K., and Kot, A. C. (2018). “Skeleton-based human action recognition with global context-aware attention LSTM networks”. *IEEE Transactions on Image Processing*, *27*(4), 1586–1599.

Liu, M., Liu, H., and Chen, C. (2018). “Robust 3D action recognition through sampling local appearances and global distributions”. *IEEE Transactions on Multimedia*, *20*(8), 1932–1947.

Liu, S., Bai, X., Fang, M., Li, L., and Hung, C.-C. (2022). “Mixed graph convolution and residual transformation network for skeleton-based action recognition”. *Applied Intelligence*, *52*(2), 1544–1555.

Liu, Y., Jiang, X., Sun, T., and Xu, K. (2019). “3D gait recognition based on a CNN-LSTM network with the fusion of skegei and da features”. In *Proceedings of the 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1–8.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). “Effective approaches to attention-based neural machine translation”. *CoRR*, *abs/1508.04025*.

Ma, G., Wu, L., and Wang, Y. (2017). “A general subspace ensemble learning framework via totally-corrective boosting and tensor-based and local patch-based extensions for gait recognition”. *Pattern Recognition*, *66*, 280 – 294.

Mahfouf, Z., Merouani, H. F., Bouchrika, I., and Harrati, N. (2018). “Investigating the use of motion-based features from optical flow for gait recognition”. *Neurocomputing*, *283*, 140–149.

Mallick, T., Das, P. P., and Majumdar, A. K. (2014). “Characterizations of noise in kinect depth images: A review”. *IEEE Sensors Journal*, *14*(6), 1731–1740.

Munaro, M., Basso, A., Fossati, A., Van Gool, L., and Menegatti, E. (2014). “3D reconstruction of freely moving persons for re-identification with a depth sensor”. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 4512–4519.

Munaro, M., Ghidoni, S., Dizmen, D. T., and Menegatti, E. (2014). “A feature-based approach to people re-identification using skeleton keypoints”. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 5644–5651.

Nambiar, A., Bernardino, A., Nascimento, J. C., and Fred, A. (2017a). “Context-aware person re-identification in the wild via fusion of gait and anthropometric features”. In *Proceedings of the 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017)*, 973–980.

Nambiar, A. M., Bernardino, A., Nascimento, J. C., and Fred, A. L. (2017b). “Towards view-point invariant person re-identification via fusion of anthropometric and gait features from kinect measurements.”. In *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications(VISIGRAPP) (5: VISAPP)*, 108–119.

Nanni, L., Munaro, M., Ghidoni, S., Menegatti, E., and Brahnam, S. (2016). “Ensemble of different approaches for a reliable person re-identification system”. *Applied Computing and Informatics*, *12*(2), 142–153.

Ng, W., Zhang, M., and Wang, T. (2022). “Multi-localized sensitive autoencoder-attention-LSTM for skeleton-based action recognition”. *IEEE Transactions on Multimedia*, *24*, 1678–1690.

Nguyen, T.-N., Pham, D.-T., Le, T.-L., Vu, H., and Tran, T.-H. (2018). “Novel skeleton-based action recognition using covariance descriptors on most informative joints”. In *Proceedings of the 10th International Conference on Knowledge and Systems Engineering (KSE)*, 50–55.

Nie, Q., Wang, J., Wang, X., and Liu, Y. (2019). “View-invariant human action recognition based on a 3D bio-constrained skeleton model”. *IEEE Transactions on Image Processing*, *28*(8), 3959–3972.

Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., and Bajcsy, R. (2014). “Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition”. *Journal of Visual Communication and Image Representation*, *25*(1), 24–38.

Oreifej, O. and Liu, Z. (2013). “HON4D: Histogram of oriented 4d normals for activity recognition from depth sequences”. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 716–723.

Papadopoulos, G. T., Axenopoulos, A., and Daras, P. (2014). “Real-time skeleton-tracking-based human action recognition using kinect data”. In Gurrin, C., Hopfgartner, F., Hurst, W., Johansen, H., Lee, H., and O’Connor, N. (Eds.), *MultiMedia Modeling*, 473–483., Cham. Springer International Publishing.

Pham, H.-H., Khoudour, L., Crouzil, A., Zegers, P., and Velastin, S. A. (2019). "Learning to recognise 3D human action from a new skeleton-based representation using deep convolutional neural networks". *IET Computer Vision*, *13*(3), 319–328.

Pickering, C. A., Burnham, K. J., and Richardson, M. J. (2007). "A research study of hand gesture recognition technologies and applications for human vehicle interaction". In *Proceedings of the 3rd Institution of Engineering and Technology Conference on Automotive Electronics*, 1–15.

Preis, J., Kessel, M., Werner, M., and Linnhoff-Popien, C. (2012). "Gait recognition with kinect". In *Proceedings of the 1st international workshop on kinect in pervasive computing*, 1–4. New Castle, UK.

Presti, L. L. and La Cascia, M. (2016). "3D skeleton-based human action classification: A survey". *Pattern Recognition*, *53*, 130–147.

Rahman, M. W. and Gavrilova, M. L. (2017). "Kinect gait skeletal joint feature-based person identification". In *Proceedings of the IEEE 16th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC)*, 423–430.

Rahmani, H., Mahmood, A., Huynh, D., and Mian, A. (2016). "Histogram of oriented principal components for cross-view action recognition". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(12), 2430–2443.

Rahmani, H., Mahmood, A., Huynh, D. Q., and Mian, A. (2014). "Real time action recognition using histograms of depth gradients and random decision forests". In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 626–633.

Rahmani, H. and Mian, A. (2016). "3D action recognition from novel viewpoints". In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1506–1515.

Ramanathan, M., Yau, W., and Teoh, E. K. (2014). "Human action recognition with video data: Research and evaluation challenges". *IEEE Transactions on Human-Machine Systems*, *44*(5), 650–663.

Ramezani, M. and Yaghmaee, F. (2016). "A review on human action analysis in videos for retrieval applications". *Artificial Intelligence Review*, *46*(4), 485–514.

Rao, H., Wang, S., Hu, X., Tan, M., Da, H., Cheng, J., and Hu, B. (2020). "Self-supervised gait encoding with locality-aware attention for person re-identification". In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 898–905.

Rao, H., Wang, S., Hu, X., Tan, M., Guo, Y., Cheng, J., Liu, X., and Hu, B. (2021). "A self-supervised gait encoding approach with locality-awareness for 3D skeleton based person re-identification". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). “You only look once: Unified, real-time object detection”. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Redmon, J. and Farhadi, A. (2018). “YOLOv3: An incremental improvement”. *CoRR, abs/1804.02767*.

Rida, I., Almaadeed, N., and Almaadeed, S. (2019). “Robust gait recognition: a comprehensive survey”. *IET Biometrics*, *8*(1), 14–28.

Romaissa, B. D., Mourad, O., and Brahim, N. (2021). “Vision-based multi-modal framework for action recognition”. In *Proceedings of the 25th International Conference on Pattern Recognition (ICPR)*, 5859–5866.

Rădulescu, B. A. and Florea, A. M. (2021). “Human action recognition methods based on CNNs for RGB video input”. In *Proceedings of the 23rd International Conference on Control Systems and Computer Science (CSCS)*, 112–118.

Saggese, A., Strisciuglio, N., Vento, M., and Petkov, N. (2019). “Learning skeleton representations for human action recognition”. *Pattern Recognition Letters*, *118*, 23–31.

Sameem, M. S. I., Bakhat, K., Khan, R., Iqbal, M., Islam, M. M., and Ye, Z. (2021). “Action recognition using interrelationships of 3D joints and frames based on angle sine relation and distance features using interrelationships”. *Applied Intelligence*, *51*, 6001–6013.

Sanchez-Caballero, A., de Lopez-Diz, S., Fuentes-Jimenez, D., Losada-Gutiérrez, C., Marrón-Romera, M., Casillas-Perez, D., and Sarker, M. I. (2022). “3DFCNN: Real-time action recognition using 3D deep neural networks with raw depth information”. *Multimedia Tools and Applications*, *81*(17), 24119–24143.

Sanchez-Caballero, A., Fuentes-Jimenez, D., and Losada-Gutiérrez, C. (2020). “Exploiting the ConvLSTM: Human action recognition using raw depth video-based recurrent neural networks”. *arXiv preprint arXiv:2006.07744*.

Savitzky, A. and Golay, M. J. (1964). “Smoothing and differentiation of data by simplified least squares procedures”. *Analytical chemistry*, *36*(8), 1627–1639.

Semwal, V. B., Singha, J., Sharma, P. K., Chauhan, A., and Behera, B. (2017). “An optimized feature selection technique based on incremental feature analysis for biometric gait data classification”. *Multimedia tools and applications*, *76*(22), 24457–24475.

Sepas-Moghaddam, A. and Etemad, A. (2023). “Deep gait recognition: A survey”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(1), 264–284.

Shahroudy, A., Liu, J., Ng, T.-T., and Wang, G. (2016). “NTU RGB+D: A large scale dataset for 3D human activity analysis”. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1010–1019.

Shao, Z., Li, Y., and Zhang, H. (2021). "Learning representations from skeletal self-similarities for cross-view action recognition". *IEEE Transactions on Circuits and Systems for Video Technology*, *31*(1), 160–174.

Shi, Z. and Kim, T.-K. (2017). "Learning and refining of privileged information-based RNNs for action recognition from depth sequences". In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3461–3470.

Sima, M., Hou, M., Zhang, X., Ding, J., and Feng, Z. (2022). "Action recognition algorithm based on skeletal joint data and adaptive time pyramid". *Signal, Image and Video Processing*, 1–8.

Singh, J. P., Jain, S., Arora, S., and Singh, U. P. (2018). "Vision-based gait recognition: A survey". *IEEE Access*, *6*, 70497–70527.

Singh, R., Khurana, R., Kushwaha, A. K. S., and Srivastava, R. (2020). "Combining CNN streams of dynamic image and depth data for action recognition". *Multimedia Systems*, 1–10.

Song, Y., Tang, J., Liu, F., and Yan, S. (2014). "Body surface context: A new robust feature for action recognition from depth videos". *IEEE Transactions on Circuits and Systems for Video Technology*, *24*(6), 952–964.

Sreela, S. and Idicula, S. M. (2018). "Action recognition in still images using residual neural network features". *Procedia Computer Science*, *143*, 563 – 569.

Sun, J., Wang, Y., Li, J., Wan, W., Cheng, D., and Zhang, H. (2018). "View-invariant gait recognition based on kinect skeleton feature". *Multimedia Tools and Applications*, *77*(19), 24909–24935.

Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., and Liu, J. (2023). "Human action recognition from various data modalities: A review". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(3), 3200–3225.

Tafazzoli, F. and Safabakhsh, R. (2010). "Model-based human gait recognition using leg and arm movements". *Engineering Applications of Artificial Intelligence*, *23*(8), 1237 – 1246.

Tang, J., Luo, J., Tjahjadi, T., and Guo, F. (2017). "Robust arbitrary-view gait recognition based on 3D partial similarity matching". *IEEE Transactions on Image Processing*, *26*(1), 7–22.

Tsironi, E., Barros, P., Weber, C., and Wermter, S. (2017). "An analysis of convolutional long short-term memory recurrent neural networks for gesture recognition". *Neurocomputing*, *268*, 76 – 86.

Tu, Z., Li, H., Zhang, D., Dauwels, J., Li, B., and Yuan, J. (2019). "Action-stage emphasized spatiotemporal vlad for video action recognition". *IEEE Transactions on Image Processing*, *28*(6), 2799–2812.

Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., and Baik, S. W. (2018). “Action recognition in video sequences using deep bi-directional LSTM with CNN features”. *IEEE Access*, *6*, 1155–1166.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). “Attention is all you need”. *Advances in neural information processing systems*, *30*.

Vemulapalli, R., Arrate, F., and Chellappa, R. (2014). “Human action recognition by representing 3D skeletons as points in a lie group”. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 588–595.

Vemulapalli, R. and Chellappa, R. (2016). “Rolling rotations for recognizing human actions from 3D skeletal data”. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4471–4479.

Verlekar, T. T., Soares, L. D., and Correia, P. L. (2018). “Gait recognition in the wild using shadow silhouettes”. *Image and Vision Computing*, *76*, 1–13.

Vision, O. O. S. C. (2019), “OpenCV -object detection”. `https://docs.opencv.org/3.4.3/df/dfb/group__imgproc__object.html`. Accessed: 12-04-2019.

Vrigkas, M., Nikou, C., and Kakadiaris, I. A. (2015). “A review of human activity recognition methods”. *Frontiers in Robotics and AI*, *2*, 28.

Wan, C., Wang, L., and Phoha, V. V. (2018). “A survey on gait recognition”. *ACM Computing Surveys (CSUR)*, *51*(5).

Wang, H., Kläser, A., Schmid, C., and Liu, C.-L. (2011). “Action recognition by dense trajectories”. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*.

Wang, H. and Wang, L. (2018). “Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection”. *IEEE Transactions on Image Processing*, *27*(9), 4382–4394.

Wang, J., Liu, Z., Wu, Y., and Yuan, J. (2012). “Mining actionlet ensemble for action recognition with depth cameras”. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1290–1297.

Wang, J., Nie, X., Xia, Y., Wu, Y., and Zhu, S.-C. (2014). “Cross-view action modeling, learning and recognition”. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2649–2656.

Wang, L., Huynh, D. Q., and Koniusz, P. (2020). “A comparative review of recent kinect-based action recognition algorithms”. *IEEE Transactions on Image Processing*, *29*, 15–28.

Wang, P., Li, W., Gao, Z., Tang, C., and Ogunbona, P. O. (2018). “Depth pooling based large-scale 3D action recognition with convolutional neural networks”. *IEEE Transactions on Multimedia*, *20*(5), 1051–1061.

Wang, P., Li, W., Gao, Z., Zhang, J., Tang, C., and Ogunbona, P. (2015). “Deep convolutional neural networks for action recognition using depth map sequences”. *arXiv preprint arXiv:1501.04686*.

Wang, P., Li, W., Gao, Z., Zhang, J., Tang, C., and Ogunbona, P. O. (2016). “Action recognition from depth maps using deep convolutional neural networks”. *IEEE Transactions on Human-Machine Systems*, *46*(4), 498–509.

Wang, P., Li, W., Li, C., and Hou, Y. (2018). “Action recognition based on joint trajectory maps with convolutional neural networks”. *Knowledge-Based Systems*, *158*, 43–53.

Wang, Y., Sun, J., Li, J., and Zhao, D. (2016). “Gait recognition based on 3D skeleton joints captured by kinect”. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 3151–3155.

Webster, J. B. and Darter, B. J. (2019). “4 - principles of normal and pathologic gait”. In J. B. Webster and D. P. Murphy (Eds.), *Atlas of Orthoses and Assistive Devices* (Fifth Edition ed.). 49 – 62. Philadelphia: Elsevier.

Weiyao, X., Muqing, W., Min, Z., Yifeng, L., Bo, L., and Ting, X. (2019). “Human action recognition using multilevel depth motion maps”. *IEEE Access*, *7*, 41811–41822.

Whitehill, J., Serpell, Z., Lin, Y., Foster, A., and Movellan, J. R. (2014). “The faces of engagement: Automatic recognition of student engagementfrom facial expressions”. *IEEE Transactions on Affective Computing*, *5*(1), 86–98.

Wint Cho, T. Z., Win, M. T., and Win, A. (2018). “Human action recognition system based on skeleton data”. In *Proceedings of the IEEE International Conference on Agents (ICA)*, 93–98.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). “CBAM: Convolutional block attention module”. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Wu, A., Zheng, W.-S., and Lai, J.-H. (2017). “Robust depth-based person reidentification”. *IEEE Transactions on Image Processing*, *26*(6), 2588–2603.

Wu, H., Ma, X., and Li, Y. (2022). “Spatiotemporal multimodal learning with 3D cnns for video action recognition”. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(3), 1250–1261.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). “Google’s neural machine translation system: Bridging the gap between human and machine translation”. *arXiv preprint arXiv:1609.08144*.

Xia, L. and Aggarwal, J. (2013). “Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera”. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2834–2841.

Xia, L., Chen, C.-C., and Aggarwal, J. K. (2012). “View invariant human action recognition using histograms of 3D joints”. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 20–27.

Xiao, Y., Chen, J., Wang, Y., Cao, Z., Tianyi Zhou, J., and Bai, X. (2019). “Action recognition for depth video using multi-view dynamic images”. *Information Sciences*, *480*, 287–304.

Xu, K., Jiang, X., and Sun, T. (2021). “Gait recognition based on local graphical skeleton descriptor with pairwise similarity network”. *IEEE Transactions on Multimedia*, 1–1.

Yan, S., Smith, J. S., and Zhang, B. (2017). “Action recognition from still images based on deep vlad spatial pyramids”. *Signal Processing: Image Communication*, *54*, 118 – 129.

Yan, S., Xiong, Y., and Lin, D. (2018). “Spatial temporal graph convolutional networks for skeleton-based action recognition”. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Yang, K., Dou, Y., Lv, S., Zhang, F., and Lv, Q. (2016). “Relative distance features for gait recognition with kinect”. *Journal of Visual Communication and Image Representation*, *39*, 209 – 217.

Yang, T., Hou, Z., Liang, J., Gu, Y., and Chao, X. (2020). “Depth sequential information entropy maps and multi-label subspace learning for human action recognition”. *IEEE Access*, *8*, 135118–135130.

Yang, X. and Tian, Y. (2014). “Super normal vector for activity recognition using depth sequences”. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 804–811.

Yang, X. and Tian, Y. (2017). “Super normal vector for human activity recognition with depth cameras”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(5), 1028–1039.

Yang, X. and Tian, Y. L. (2012). “Eigenjoints-based action recognition using naïve-bayes-nearest-neighbor”. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 14–19.

Yang, X., Zhang, C., and Tian, Y. (2012). “Recognizing actions using depth motion maps-based histograms of oriented gradients”. In *Proceedings of the 20th ACM International Conference on Multimedia*, MM ’12, 1057–1060.

Yang, Z., Li, Y., Yang, J., and Luo, J. (2019). “Action recognition with spatio–temporal visual attention on skeleton image sequences”. *IEEE Transactions on Circuits and Systems for Video Technology*, *29*(8), 2405–2415.

Yao, B. and Fei-Fei, L. (2010). “Grouplet: A structured image representation for recognizing human and object interactions”. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 9–16.

Ye, M., Yang, C., Stankovic, V., Stankovic, L., and Cheng, S. (2020). “Distinct feature extraction for video-based gait phase classification”. *IEEE Transactions on Multimedia*, *22*(5), 1113–1125.

Yoo, J., Hwang, D., Moon, K., and Nixon, M. S. (2008). “Automated human recognition by gait using neural network”. In *Proceedings of the First Workshops on Image Processing Theory, Tools and Applications*, 1–6.

Yun, K., Honorio, J., Chattopadhyay, D., Berg, T. L., and Samaras, D. (2012). “Two-person interaction detection using body-pose features and multiple instance learning”. In *Proceedings of the IEEE computer society conference on computer vision and pattern recognition workshops (CVPRW)*, 28–35.

Zhang, B., Yang, Y., Chen, C., Yang, L., Han, J., and Shao, L. (2017). “Action recognition using 3D histograms of texture and a multi-class boosting classifier”. *IEEE Transactions on Image Processing*, *26*(10), 4648–4660.

Zhang, C., Tian, Y., Guo, X., and Liu, J. (2018). “DAAL: Deep activation-based attribute learning for action recognition in depth videos”. *Computer Vision and Image Understanding*, *167*, 37–49.

Zhang, H., Li, S., Shi, Y., and Yang, J. (2019). “Graph fusion for finger multimodal biometrics”. *IEEE Access*, *7*, 28607–28615.

Zhang, J., Li, W., Ogunbona, P. O., Wang, P., and Tang, C. (2016). “RGB-D-based action recognition datasets: A survey”. *Pattern Recognition*, *60*, 86–105.

Zhang, S., Yang, Y., Xiao, J., Liu, X., Yang, Y., Xie, D., and Zhuang, Y. (2018). “Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks”. *IEEE Transactions on Multimedia*, *20*(9), 2330–2343.

Zhang, Y., Cheng, L., Wu, J., Cai, J., Do, M. N., and Lu, J. (2016). “Action recognition in still images with minimum annotation efforts”. *IEEE Transactions on Image Processing*, *25*(11), 5479–5490.

Zhang, Z. (2012). “Microsoft kinect sensor and its effect”. *IEEE MultiMedia*, *19*, 4–12.

Zheng, B., Chen, L., Wu, M., Pedrycz, W., and Hirota, K. (2022). “Skeleton-based action recognition using two-stream graph convolutional network with pose refinement”. In *Proceedings of the 41st Chinese Control Conference (CCC)*, 6353–6356.

Zheng, L., Zha, Y., Kong, D., Yang, H., and Zhang, Y. (2022). “Multi-branch angle aware spatial temporal graph convolutional neural network for model-based gait recognition”. *IET Cyber-Systems and Robotics*, *4*(2), 97–106.

Ziaeefard, M. and Bergevin, R. (2015). “Semantic human activity recognition: A literature review”. *Pattern Recognition*, *48*(8), 2329 – 2345.

Publications

Journal Papers

1. **Rashmi M.**, Ashwin T. S. and Ram Mohana Reddy Guddeti. (2021). “Surveillance video analysis for student action recognition and localization inside computer laboratories of a smart campus”. *Multimedia Tools and Applications, 80(2)*, pp. 2907–2929. (Springer).
   DOI=https://doi.org/10.1007/s11042-020-09741-5 [SCOPUS and SCIE indexed]

2. **Rashmi M.** and Ram Mohana Reddy Guddeti. (2022). “Human identification system using 3D skeleton-based gait features and LSTM model”. *Journal of Visual Communication and Image Representation, 82*, 103416 (Elsevier).
   DOI=https://doi.org/10.1016/ j.jvcir.2021.103416 [SCOPUS and SCIE indexed]

3. **Rashmi M.** and Ram Mohana Reddy Guddeti. “Human action recognition using multi-stream attention-based deep networks with heterogeneous data from overlapping sub-actions”. *Neural Computing and Applications*. (Springer) [Under review]

4. **Rashmi M.** and Ram Mohana Reddy Guddeti. “Exploiting skeleton-based gait events with attention-guided residual deep learning model for human identification”. *Applied Intelligence*. (Springer) [Revision submitted]

Conference Papers

1. **Rashmi M.** and Ram Mohana Reddy Guddeti. (2020). “Skeleton based human action recognition for smart city application using deep learning”. In *Proceedings of the International Conference on COMmunication Systems and NETworkS (COMSNETS),* pp. 756-761. (IEEE) [Held at Bangalore, India, during Jan. 7-11, 2020]
   DOI=10.1109/COMSNETS48256.2020.9027469 [SCOPUS indexed]

2. **Rashmi M** and Ram Mohana Reddy Guddeti. (2022), “Multi-stream multi-attention deep neural network for context-aware human action recognition”. In *Proceedings of the IEEE Region 10 Symposium (TENSYMP),* pp. 1-6. [Held at IIT Bombay, Mumbai, India, during 1-3 July 2022]
   DOI: 10.1109/TENSYMP54529.2022.9864404 [SCOPUS indexed]

Other Publications not Related to Thesis

1. Sampat Kumar Ghosh, **Rashmi M.**, Biju R. Mohan, and Ram Mohana Reddy Guddeti. (2022). "Skeleton-based human action recognition using motion and orientation of joints". *In: Gupta, D., Sambyo, K., Prasad, M., Agarwal, S. (eds) Advanced Machine Intelligence and Signal Processing. Lecture Notes in Electrical Engineering, vol 858.* (Springer)
   DOI: https://doi.org/10.1007/978-981-19-0840-8_6 [SCOPUS indexed]
2. Sampat Kumar Ghosh, **Rashmi M**, Biju R. Mohan, and Ram Mohana Reddy Guddeti. (2023). "Deep learning-based multi-view 3D-human action recognition using skeleton and depth data". *Multimedia Tools and Applications, 82,* pp. 19829–19851. (Springer)
   DOI:https://doi.org/10.1007/s11042-022-14214-y [SCOPUS and SCIE indexed]

Curriculum Vitae

Mrs. Rashmi M
Full-Time Research Scholar
Department of Information Technology
National Institute of Technology Karnataka Surathkal
Srinivasanagar Post- 575 025
Karnataka State.

Permanent Address

Rashmi M
D.No. 16-66/3(10),
“Vrushanka”
Udayanagar
Near NITK, Surathkal
Srinivasnagar Post-575 025
Dakshina Kannada
Karnataka State
Email: *nm.rashmi@gmail.com*
Mobile: +919449773711.

Academic Records

1. M.Tech. in Computer Science and Engineering from NMAMIT NITTE, Karnataka State, 2014.
2. B.E. in Computer Science and Engineering from KVGCE, Sullia, Karnataka State, 2004.

Research Interests

Deep Learning
Computer Vision
Human Action Recognition
Gesture Recognition
Gait Recognition
Vision-based Smart Applications

Programming Languages

C, C++, Python, MATLAB, JavaScript, SQL.