CERTIFICATE

It is certified that the work contained in this thesis entitled "*DESIGN AND DEVELOPMENT OF DEEP LEARNING BASED APPROACHES FOR 2D and 3D HUMAN POSE ESTIMATION AND ACTIVITY RECOGNITION*" by "*Ms. PRATISHTHA VERMA (Roll No. 17071507)*" has been carried out under my supervision and that it has not been submitted elsewhere for a degree.

It is further certified that the student has fulfilled all the requirements of Comprehensive, Candidacy and SOTA.

Prof. Rajeev Srivastava

Professor

Department of Computer Science and Engineering
Indian Institute of Technology
(Banaras Hindu University)
Varanasi-221005

DECLARATION BY THE CANDIDATE

I, PRATISHTHA VERMA , certify that the work embodied in this thesis is my own bonafide work and carried out by me under the supervision of Prof. Rajeev Srivastava from DEC-2017 to JULY-2021, at the DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, Indian Institute of Technology (BHU) Varanasi. The matter embodied in this thesis has not been submitted for the award of any other degree/diploma. I declare that I have faithfully acknowledged and given credits to the research workers wherever their works have been cited in my work in this thesis. I further declare that I have not willfully lifted up any other's work, paragraphs, text, data, results, etc., reported in journals, books, magazines, reports dissertations, theses, etc., or available at websites and included them in this thesis and cited as my own work.

Pratishtha.

Date : 14-07-2021

Place : Varanasi

Signature of the Student

(PRATISHTHA VERMA)

CERTIFICATE BY THE SUPERVISOR

It is certified that the above statement made by the student is correct to the best of my/our knowledge.

14.7.2021

Signature of Supervisor

(Prof. Rajeev Srivastava)

Department of Computer Science and Engineering
Indian Institute of Technology
(Banaras Hindu University)
Varanasi-221005

आचार्य /Professor
संगणक विज्ञान एवं अभियांत्रिकी विभाग /Department of Computer Sc. & Engg.
भारतीय प्रौद्योगिकी संस्थान /Indian Institute of Technology
(बनारस हिन्दू यूनिवर्सिटी)/(Banaras Hindu University)
वाराणसी /Varanasi-221005

Sk Singh

Signature of Head of Department

(Prof. S. K. Singh)

आचार्य व विभागाध्यक्ष
Professor & Head
संगणक विज्ञान एवं अभियांत्रिकी विभाग
Department of Computer Sc. & Engg.
भारतीय प्रौद्योगिकी संस्थान
Indian Institute of Technology
(बनारस हिन्दू यूनिवर्सिटी)
(Banaras Hindu University)
वाराणसी–२२१००५ / Varanasi-221005

COPYRIGHT TRANSFER CERTIFICATE

Title of the Thesis: DESIGN AND DEVELOPMENT OF DEEP LEARNING BASED APPROACHES FOR 2D and 3D HUMAN POSE ESTIMATION AND ACTIVITY RECOGNITION

Name of the Student: PRATISHTHA VERMA

Copyright Transfer

The undersigned hereby assigns to the Indian Institute of Technology (Banaras Hindu University) Varanasi, all rights under copyright that may exist in and for the above thesis submitted for the award of the DOCTOR OF PHILOSOPHY.

Pratishtha.

Date : 14-07-2021

Place : Varanasi

(PRATISHTHA VERMA)

Note: However, the author may reproduce or authorize others to reproduce material extracted verbatim from the thesis or derivative of the thesis for author's personal use provided that the source and the Institute's copyright notice are indicated.

To

My Beloved Parents

Mrs. Rekha Verma

and

Dr. Deepak Kumar

ACKNOWLEDGEMENT

I want to take this opportunity to express my deep sense of gratitude to all who helped me directly or indirectly during this thesis work. Firstly, I would like to thank my supervisor, Prof. Rajeev Srivastava, for being a great mentor and the best adviser I could ever have. His advice, encouragement, and critics are a source of innovative ideas, inspiration, and causes behind the successful completion of this Thesis work. The confidence shown on me by him was the most significant source of inspiration for me. It has been a privilege working with him for several years. I am highly obliged to all the faculty members of the Computer Science and Engineering Department for their support and encouragement. I express my sincere thanks to Prof. A. K. Tripathi, Prof. K. K. Shukla, Prof. S.K.Singh, and Dr. Prateek Chattopadhyay of the Department of Computer Science and Engineering and Prof. Subir Das, Department of Mathematical Sciences IIT (BHU), for providing continuous support, encouragement, and advice. I express my sincere thanks to all the Professors, Deans, office staff, supporting staff, and Ph.D. Research Scholars of Indian Institute of Technology (BHU) Varanasi India. I express my gratitude to the Director, Registrars, Deans, Heads, and Student Alumni of the Indian Institute of Technology (BHU) Varanasi.

My memory of the study period at IIT (BHU) can never be complete without mentioning my fellow research scholars. Special thanks to Dr. Vibhav Prakash Singh, Dr. Roshan Singh, Dr. Gargi Srivastava, Mr. Ankit Jaiswal, Dr. Kuntesh Ketan Jani, Mr. Surya Kant Singh, Mr. Ashish Kumar Sharma, Mrs. Divya Singh and Mr. Santosh Kumar Tripathy for their great help and cooperation.

I extend special thanks to the non-teaching staff in the department, particularly the late Mr. A. N. Yadav, Mr. Manoj Kumar Singh, Mr. Ravi Kumar Bharti, Mr. Prakhar Kumar, Mr. Ritesh Singh, and Mr. Shubham Pandey for their consistent support.

My parents, Dr. Deepak Kumar and Smt. Rekha Verma, who gave me the power and brain to work out on this research and their help at every level made me to see this success. I owe thanks to my elder sister, Ms. Akanksha Verma for her continued and unfailing love, support and understanding during my pursuit of Ph.D degree that made the completion of thesis possible. You were always around at times I thought that it is impossible to continue, you helped me to keep things in perspective. I greatly value her contribution and deeply appreciate her belief in me. I extend my thanks to my brother-in-laws Mr. Chander Soni who is my part of inspiration. Finally, I would like to wind up by paying my heartfelt thanks and prayers to the Almighty for his unbound love and grace.

- Pratishtha Verma

Contents

List of Figures **xv**

List of Tables **xvii**

Abbreviations **xix**

Symbols **xxi**

Preface **xxiii**

1 INTRODUCTION **1**

1.1 Background 1

1.2 Motivation 3

1.3 Problem Statement 6

1.4 Thesis Objective 7

1.5 Outline of the thesis 10

2 THEORETICAL FOUNDATION AND LITERATURE SURVEY **15**

2.1 Introduction 16

2.2 Applications 20

2.2.1 Human-Computer Interaction 20

2.2.2 Video Surveillance 22

2.3 Conventional and deep learning based human pose estimation techniques 22

2.3.1 Analysis 24

2.3.2 Input Modality 25

2.3.2.1 RGB Image 25

2.3.2.2 Depth Image 25

2.3.2.3 Infra-red(IR) Image 26

2.3.3 Conventional Approach 26

2.3.3.1 Preprocessing 26

Information calibration 26

Localization 27

Human Detection 27
Background Substraction 27
2.3.3.2 Feature Discription 28
Low-level features: 30
High-level features: 30
Motion features: 30
2.3.3.3 Body Modelling 32
2.3.3.4 Methodology 34
Generative and Descriminative Methods: 35
I. Descriminative methods: 35
(i) Mapping and Learning based methods: 36
(ii) Deep laerning based HPE methods: 37
Direct prediction: 37
Indirect prediction: 38
Single and Multi-stage approaches: 38
II. Generative based methods: 40
2.3.4 Post-processing: 41
2.4 Deep learning based human activity recognition techniques: 42
2.4.1 Convolutional neural network 43
2.4.2 Autoencoder 43
2.4.3 Recurrent neural network 43
2.5 Benchmark and Evaluation metrics: 44
2.5.1 For human pose estimation method: 44
2.5.1.1 Evaluation metric: 45
2.5.2 Evaluation metrics and datasets for HAR 47
Cornell activity datasets (CAD-60) 47
CAD-120 47
UTD-MHAD 48
2.6 Comparative analysis of recent state-of-the-art techniques for human pose estimation. 48
2.7 Challenges 49
2.7.1 2D Human Pose Estimation 49
2.7.2 3D Human Pose Estimation 52
2.7.3 Human Activity Recognition 53
2.8 Conclusion 54

3 THE TWO DEEP NETWORKS FOR 2D HUMAN POSE ESTIMATION 55
3.1 Introduction 56
3.2 Literature 61
3.2.1 Multi-stage deep model for 2D human pose estimation 61
Convolution layer 62
Feature mapping layer 63

Pooling Layer 63
3.2.2 2D human pose estimation using detection followed by estimation approach 64
3.3 Methods 64
3.3.1 Multi-stage Deep learning network for 2D HPE 65
3.3.1.1 Feature Extraction 68
3.3.1.2 Feature Refinement 69
3.3.1.3 Fusion 70
3.3.2 Detection followed by estimation Deep learning model for 2D HPE 73
3.3.2.1 Detection module 73
3.3.2.2 Feature Extraction and Classification 73
3.3.2.3 Pre trained CNN based model for making detection accurate 76
3.3.2.4 Deep architecture for HPE 77
3.4 Result and Discussion 78
3.4.1 Multi-stage Deep learning network for 2D HPE 78
3.4.1.1 MPII Dataset 78
3.4.1.2 LSP Dataset 78
3.4.1.3 Training Details 79
3.4.1.4 Evaluation Metric Used 79
3.4.1.5 Preprocessing 81
3.4.1.6 Experiments 82
Experiment 1: 83
Experiment 2: 83
3.4.2 Detection followed by estimation Deep learning model for 2D HPE 84
3.4.2.1 Experiments 84
INIRIA Person dataset: 84
MPII and LSP Dataset: 85
3.5 Conclusion 86

4 DEEP LEARNING APPROACHES FOR 3D HUMAN POSE ESTIMATION FROM SINGLE VIEW PERSPECTIVE 87
4.1 Introduction 88
4.2 Literature 91
4.3 Methods 95
4.3.1 Two-stage deep network for 3D human pose estimation by exploiting spatial data via its 2D pose 95
4.3.2 Three stage deep network for 3D human pose estimation by exploting spatial and temporal data 97
4.3.2.1 Frame specific pose estimation(FSPE) 2 97
4.3.2.2 $Spatial_{3D}Reconstruction$ 101
4.3.2.3 Feature Residual connection(FRC) strategy 102

4.3.2.4 $Temporal_{3D}Reconstruction$ 104

4.4 Experiments and Results 105

4.4.1 Two-stage deep network for 3D human pose estimation by exploiting spatial data via its 2D pose 105

4.4.1.1 Dataset and Metrics: 105

MPII Dataset: 105

Human3.6M dataset: 105

HumanEva-I dataset: 106

Evaluation Metric: 106

4.4.1.2 Implementation Details: 107

4.4.1.3 Training Details 107

4.4.1.4 Results and Ablation Analysis: 108

4.4.1.5 MPII Dataset: 108

4.4.1.6 Human3.6M Dataset: 109

4.4.1.7 HumanEva-I 109

4.4.2 Three stage deep network for 3D human pose estimation by exploting spatial and temporal data 111

4.4.2.1 Dataset and evaluation protocols for 2D pose estimation 111

MPII Dataset. 112

Evaluation Protocol(PCKh) 112

4.4.2.2 Dataset and evaluation protocols for 3D pose estimation 113

Human3.6M Dataset. 113

HumanEva-I. 113

Evaluation Protocols. 114

4.4.2.3 Implementation Details: 115

Training Details. 115

4.4.2.4 Result and Discussion 116

4.4.2.5 Results on MPII 116

4.4.2.6 Results on Human3.6M 117

4.4.2.7 Results on HumanEva-I 118

4.4.2.8 Failure results 118

4.5 Conclusion 118

5 DEEP NETWORK FOR 3D HUMAN POSE ESTIMATION USING MULTI-VIEW DATA 123

5.1 Introduction 123

5.2 Related Work 127

5.3 Proposed Method 130

5.3.1 Enhanced stack-hourglass approach (ESHA) 130

5.4 Multi-view spatial 3D reconstruction approach 132

Convolutional(C)-ReLU(R) layers 133

Batch Normalization(BN) and Dropout(D) 135

Residual Module 136
Image/2D joint heatmaps and its fusion 136
Early Fusion 137
Late Fusion 137
5.5 Experiments 139
5.5.1 Dataset and Evaluation protocol for ESHA 139
MPII Dataset. 139
Evaluation Metric(PCKh) 139
5.5.2 Datasets and Evaluation protocol for 3D HPR 140
Evaluation Metric for 3D HPR 140
Human3.6M Dataset. 141
Evaluation-protocol#1 141
5.5.3 Implementation Details 141
5.5.4 Results and Discussions 142
5.5.4.1 Results and comparison with state-of-the-art 143
MPII Dataset. 143
Human3.6M Dataset. 143
5.5.4.2 False Results 144
5.6 Conclusion 145
5.7 conclusion 145

6 ACTIVITY RECOGNITION USING RGB AND SKELETON DATA INPUT 151
6.1 INTRODUCTION 151
6.2 Related Work 155
6.3 Methodology 157
6.3.1 RGB frames 157
6.3.1.1 Motion History Image (MHI) and Motion Energy Image (MEI) 158
6.3.1.2 First Stream 160
6.3.2 Skeleton Joints 162
6.3.2.1 From Skeleton sequences directed to the Images 163
6.3.2.2 Second Stream 168
6.3.3 Cyclic Learning Rate 172
6.3.4 Decision Level Fusion(DLF) approach 173
6.3.4.1 Weighted Product Model (WPM) 173
6.4 Experiments 174
6.4.1 UTD-MHAD Dataset 175
6.4.2 CAD-60 176
6.4.3 NTU-RGB+D120 Dataset 177
6.5 Conclusion and Future scope 179

7 CONCLUSION AND FUTURE WORK 183

7.1 Conclusions 183
7.2 Future Work 184

References 187

List of Publications 219

List of Figures

2.1 Number of research publications in the area of Human Pose Estimation (HPE) and Human Activity Recognition (HAR). 17

2.2 Some applications for HPE and HAR (from top to bottom and left to right): Human-robot interaction, human-computer interaction, gaming, video surveillance, sports, and proxemics. 17

2.3 A brief overview of the related tasks with Human Pose Estimation (HPE) 18

2.4 Goals of Preprocessing Techniques 29

2.5 The framework of the review which mainly contains four stages (left to right): Pre-processing, Feature Representation, human body models and Methodology. 29

2.6 Image features 32

2.7 Direct method. 39

2.8 Indirect method. 39

2.9 Convolutional neural network architecture. 40

2.10 False human pose. 42

3.1 The proposed architecture of first model 65

3.2 First stream of feature extraction. 66

3.3 Second stream of feature extraction. 67

3.4 Cascaded feature integration architecture. 70

3.5 The architecture of Detection module. 73

3.6 The multi-layer architecture with Fully Convolutional GoogleNet. 77

3.7 The visualization of heatmap for all 16 joint locations on the MPII dataset. The images from left to right and top to bottom have been added the heatmaps for pelvis, thorax, upper neck, head top, right shoulder, right elbow, right wrist, left shoulder, left elbow, left wrist, right knee, right ankle, left hip, left knee, left ankle, right hip. 81

3.8 The visualization of the Qualitative evaluations on both MPII dataset. . . 82

4.1 The framework of the proposed method. 96

4.2 First stage of the network. 96

4.3 The overall pipeline of the proposed method 98

4.4 Feature Extraction for FSPE Module. 101

4.5 Feature Refinement with MSCFC Strategy. 101

4.6 Visualization of the image and its 3D reconstruction. 112
4.7 The MPJPE value on FSPE based (indicated by blue color) and 2D GT based (indicated by orange color) 3D HPR method having activity order mentioned in Table 4.2 119
4.8 The P-MPJPE value on FSPE based (indicated by blue color) and 2D GT based (indicated by orange color) 3D HPR method having activity order mentioned in Table 4.2 119
4.9 The MPJPE error value for different value of NSL =5, 10, and 20 with respective LSTM layers. 119
4.10 Few of the erraneous poses of the SittingDown Class 121

5.1 The framework of the proposed method. 129
5.2 The heatmaps for different joints. 133
5.3 Early and Late fusion network over Heatmap and Frame Flow 134
5.4 Multi-view images and its 3D pose. 146
5.5 Failure Cases 146
5.6 Illumination change output 146
5.7 The influence of change of image resolution. 147
5.8 Early and Late fusion network over Heatmap and Frame Flow 147
5.9 Comparision chart for reconstruction error 147
5.10 Early and Late fusion network over Heatmap and Frame Flow 148

6.1 Few RGB and skeleton sequences for the activity ‘Right arm swipe to the left’ from UTD-MHAD dataset. 155
6.2 Proposed framework. 158
6.3 Architecture of the multi-input convnet. Here the features extracted from the MEI are combined with the features extracted from the MHI. This network has been inspired by the VGG-16 model but in order to prevent overfitting, the number of elements of the layers has been reduced. 160
6.4 Stacked histogram of joint contribution for 2 actions in UTD-MHAD dataset: 1. Forward Lunge 2. SQUAT. Each bar indicates the contribution ratio of each joint to the action in one instance. 162
6.5 Sample images of MHI, MEI, Skeleton Images(Front View, Side View, and Top View) created RGB and skeletal data of UTD-MHAD dataset for activities 1. Draw X, 2. Boxing, 3. Tennis Swing 169
6.6 Architecture of the convnet model with supervision layers used for testing/training of a single view of the skeleton image. 171
6.7 The overall network with the fusion of 3-convnet streams of skeleton image. Intermediate supervision is present to avoid overfitting. 171

List of Tables

2.1 Comparative analysis of state-of-the-art techniques on the basis of type of preprocessing, feature description, body model, method, dataset, and metric they use. 49
2.2 List and Details of largely utilized datasets for HPE. 50
2.3 Comprative analysis of state-of-the-art for 2D human pose estimation . . 51
2.4 Comprative analysis for 3D human pose estimation 52

3.1 **The result of the first introduced technique and its comparison with other state of the art on MPII dataset using modified PCK called PCKh with @0.5.** 80
3.2 **The result of first introduced technique and comparison with other state of the art on LSP dataset using PCK with @0.2.** 80
3.3 **Some Evaluated Metrics on INIRIA Person Dataset** 84
3.4 PCKh metric at 0.5 for MPII Dataset. 85
3.5 PCKh metric at 0.2 for LSP Dataset. 85

4.1 The outcome of proposed approach and its comparison with published state-of-the-art techniques on the basis of PCK0.5 metric. 110
4.2 The outcome of proposed approach and its comparison with published state-of-the-art techniques on the basis of MPJPE metric. 110
4.3 The outcome of proposed approach and its comparison with published state-of-the-art techniques on the basis of P-MPJPE metric. 111
4.4 The result of proposed method and comparison with other state-of-the-art using MPJPE metric on HumanEva-I dataset. 111
4.5 The result of FSPE method and comparison with other state-of-the-art using PCKh with @0.5. 113
4.6 The Human3.6M dataset has been utilized for 3D pose reconstruction errors(in millimeters) based on Evaluation Criteria 1. 120
4.7 The Human3.6M dataset has been utilized for 3D pose reconstruction errors(in millimeters) based on Evaluation Criteria 2. 121
4.8 The Human3.6M dataset has been utilized for 3D pose reconstruction errors(in millimeters) based on Evaluation Criteria 3. 122
4.9 The HumanEva-I dataset has been utilized for 3D pose reconstruction errors(in millimeters) based on Evaluation Criteria 1. 122

5.1 The ESHA algorithm result and its comparability with distinct state-of-the-art procedures for the basis evaluation metric PCKh with @0.5. . . . 140
5.2 The Evaluation Criteria 1 over Human3.6M public data set has been employed to calculate the 3D HPR errors (in the form of millimeters). 148
5.3 The Evaluation Criteria 2 over Human3.6M public data set has been employed to calculate the 3D HPR errors (in the form of millimeters). 149
5.4 The Evaluation Criteria 3 over Human3.6M public data set has been employed to calculate the 3D HPR errors (in the form of millimeters). 149

6.1 UTD-MHAD dataset evaluation for activity recognition techniques . . . 178
6.2 Performance of individual streams and complete system on UTD-MHAD 178
6.3 Activity Recognition methods on CAD-60 179
6.4 NTU-RGB+D120 dataset evaluation for activity recognition techniques . 179
6.5 Confusion Matrix of 27 classes of UTD-MHAD 180
6.6 Confusion Matrix of 12 classes of CAD-60. 181

Abbreviations

CV	**C**omputer **V**ision
HPE	**H**uman **P**ose **E**stimation
HAR	**H**uman **A**ctivity **R**ecognition
CNN	**C**onvolutional **N**eural **N**etwork
HOG	**H**istogram of **O**riented **G**radient
LBP	**L**ocal **B**inary **P**atterns
SVM	**S**upport **V**ector **M**achine
ReLU	**Re**ctified **L**inear **U**nit
LSTM	**L**ong **S**hort **T**erm **M**emory
MPJPE	**M**ean **P**er **J**oint **E**rror
P-MPJPE	**P** **M**ean **P**er **J**oint **E**rror
RNN	**R**ecurrent **N**eural **N**etwork
MHI	**M**otion **H**istory **I**mage
MEI	**M**otion **E**energy **I**mage
DOF	**D**egree **O**f **F**reedom
MoCap	**Mo**tion **Cap**ture
IMU	**I**nertial **M**easurement **U**nits
FSPE	**F**rame **S**pecific **P**ose **E**stimation
MSCFC	**M**ulti **S**tage **C**ascaded **C**onnection
FRC	**F**eature **R**esidual **C**onnection
AR	**A**ugmented **R**eality
AI	**A**rtificial **I**ntelligence

UAV	**U**nmanned **P**er **J**oint **E**rror
MLP	**M**ulti **L**ayer **P**erceptron
HBM	**H**uman **B**ody **M**odel
KM	**K**inematic **M**odels
BME	**B**ayesian **M**ixture of **experts**
PCP	**P**ercentage of **C**orrectly estimated **P**arts
PCK	**P**ercentage **C**orrect **K**eypoints
PDJ	**P**ercentage **D**etected **J**oints
DCDN	**D**ensely **C**onnected **D**eep **N**etwork
ESHA	**E**nhanced **S**tack **H**ourglass **A**pproach
FE	**F**eature **E**xtraction
FR	**F**eature **R**efinement
BN	**B**atch **N**ormalization
D	**D**ropout
IJS	**I**nformative **J**oints **S**core
DJ	**D**istance **J**oint

Symbols

$\varepsilon(u, u')$	3D Error
$m_i(u)$	Predicted joint
$m_i(u')$	Ground-truth joint
$m_{root}(u')$	predicted root joint
$m_{root}(u)$	root joint of the ground-truth
y_e	ground truth 3D coordinate of end points of part e
z_e	ground truth 3D coordinate of start points of part e
z_i	predicted pose
$z_{i'}$	ground-truth pose
σ	non linear activation function
x_i	represents the input of the convolutional layer
k_{ij}	represents the parameter for the convolutional kernel
b_i	bias
sl	step length for the convolution
Y_x	2D body joint coordinate
P_x	Probability of that joint
NO_J	Number of joints
H_p	Feature map of body part
H_c	Context map
$F_1(I)$	First stream feature
$F_2(I)$	Second stream feature
K_s^P	Soft-argmax operation

K_s^P	Soft-argmax operation
$M_\Gamma(m, n, t)$	represents the MHI
$J_i(p)$	coordinate of the i^{th} joint in the p^{th} frame
$D_i(p)$	distance between the vectors $J_i(p)$ and $J_0(1)$

PREFACE

Due to technological advancement, massive multimedia data is usually available in image and video forms. Many demanding applications currently use pictures and videos like human-computer interaction, automatic surveillance systems, 3D scene understanding, sports performance analysis, etc. This thesis addresses Human Pose Estimation (HPE) and Human Activity Recognition (HAR) problem. These tasks are highly utilized in the above-mentioned applications. The HPE problem is divided into two categories based on the application requirement. The first one is 2D HPE, and the second is 3D HPE. Many state-of-the-art techniques have already been introduced for HPE and HAR problem solutions, but there are many challenges yet to be solved.

This thesis first presents a detailed survey for 2D HPE, 3D HPE, and HAR problems, followed by research gaps and challenges. Second, this thesis addresses the problem of 2D HPE from the single-view perspective. The task has further been classified into a direct 2D HPE and detection-based 2D HPE.

For detection-based 2D HPE, a machine and deep learning-based model has been proposed that detects the region of interest having a human as a subject and then estimates the respective body joints' locations. For the detection part, the HOG and LBP features have been extracted and then fused. The resultant fused features have been utilized with lib-SVM for classification purposes. A CNN-based deep model has been used to make the detected proposals more accurate. Finally, a deep network has been proposed for estimation of the joint locations over the detected part. The model has been examined on publically available datasets like the INIRIA person dataset, LSP, and MPII pose estimation dataset. The proposed method gives a remarkable performance on these datasets as compared to other state-of-the-art methods.

The direct 2D HPE method directly estimates the 2D body joint coordinates from an input image. The model consists of three consecutive parts: first, a two-branch input DCNN network has been used for feature extraction. Second, a cascaded feature integration technique has been used for feature refinement that is based on the stacked hourglass method. Finally, the fusion has been performed over the detected part heatmap and the context heatmap to make the system more accurate towards occlusion. To check the effectiveness of the model, the evaluation has been performed on two standard datasets MPII and LSP. The proposed method improves pose prediction results on the PCK metric and has been compared to other state-of-the-art techniques.

Later on in this thesis, we have extended our research towards estimating 3D human pose, which includes single view and multi-view data inputs.

For 3D HPE, three models have been proposed, where the first two work on single view data input and the third works on multi-view data input. The first model uses spatial data information for 3D HPR. For this, a two-stage deep learning model has been introduced. In the first stage, the model estimates the 2D joint key points, and in the next step, a densely connected deep architecture has been used for 3D HPE.

The second model over a single view data input uses spatial and temporal information to estimate 3D human pose. In this method, a three-stage deep learning model has been proposed. First, the model evaluates the 2D Human pose. Then, in the second stage, basic deep learning concepts like convolution, batch normalization, ReLU, and dropout have been utilized for spatial 3D pose estimation. In the last step, LSTM deep architecture has been used for temporal refinement. The third model is from multi-view data input. This model proposes a two-stage multi-view 3D HPR technique using deep learning that utilizes both the direct and 2D pose information in a multi-view scenario to estimate the 3D pose. All the proposed 3D HPE techniques give outstanding results on MPJPE, P-MPJPE, and N-MPJPE metrics over the Human3.6M dataset compared to other state-of-the-art approaches.

Finally, in this thesis, a conventional and machine learning-based technique has been proposed that uses RGB data with its 3D skeleton information for HAR. The model takes help of a convolutional neural network (Convnet) and long short-term memory (LSTM) as a recurrent neural network (RNN). The proposed method has two parts: first, motion representation images like motion history images (MHI) and motion energy images (MEI) have been created from the RGB videos. The convnet has been trained, using these images with feature-level fusion. Second, the skeleton data have been utilized with a proposed algorithm that develops skeleton intensity images for three views (top, front, and side). Each view is first analyzed by a convnet, which generates the set of feature maps fused for further analysis. On top of convnet sub-networks, LSTM has been used to exploit the temporal dependency. The softmax scores from these two independent parts are later combined at the decision level—the suggested approach privileges the perfect utilization of RGB and skeleton data available from an RGB-D sensor. The proposed approach has been tested on three famous and challenging multimodal datasets: UTD-MHAD, CAD-60, and NTU-RGB + D120. Results have shown the method gives a satisfactory outcome as compared to the other state-of-the-art.