

HUMAN ACTION AND ACTIVITY RECOGNITION USING VIDEO SEQUENCES

A thesis Submitted to

DELHI TECHNOLOGICAL UNIVERSITY

For the Award of degree of

DOCTOR OF PHILOSOPHY

In

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

by

TEJ SINGH

(2K16/PhD/EC/05)

Under the Supervision of

Dr. Dinesh Kumar Vishwakarma

Associate Professor, Department of Information Technology, DTU



Department of Electronics & Communication Engineering

Delhi Technological University

(Formerly Delhi College of Engineering)

AUGUST 2020



© DELHI TECHNOLOGICAL UNIVERSITY-2020
ALL RIGHTS RESERVED

DECLARATION

I declare that the research work reported in the thesis entitled “**HUMAN ACTION and ACTIVITY RECOGNITION using VIDEO SEQUENCES**” for the award of the degree of *Doctor of Philosophy* in the *Department of Electronics and Communication Engineering* has been carried out by me under the supervision of Dr. *Dinesh Kumar Vishwakarma*, Associate Professor in Department of Information and Technology, Delhi Technological University, Delhi, India.

The research work embodied in this thesis, except where otherwise indicated, is my original research. This thesis has not been submitted by earlier in part or full to any other University or Institute for the award of any degree or diploma. This thesis does not contain other person’s data, graphs or other information unless specifically acknowledged.

Date:

Tej Singh

2K16/Ph.D./EC/05

CERTIFICATE

This is to certify that the work contained in the thesis entitled “**HUMAN ACTION and ACTIVITY RECOGNITION using VIDEO SEQUENCES**” submitted by Mr. Tej Singh (**Reg. No.: 2K16/Ph.D./EC/05**) for the award of degree of Doctor of Philosophy to the Delhi Technological University is based on the original research work carried out by him. He has worked under my supervision and has fulfilled the requirements as per the requisite standard for the submission of the thesis. It is further certified that the work embodied in this thesis has neither partially nor fully submitted to any other university or institution for the award of any degree or diploma.

Dr. Dinesh Kumar Vishwakarma
Supervisor, Associate Professor
Department of Information Technology
Delhi Technological University, Delhi

ACKNOWLEDGMENT

I am indebted to God under whose power I pursued this Ph. D. Thanks to Almighty for granting me wisdom health and strength to undertake this research task and enabling me to its completion. I pray that I can always serve him in whatever he wants me to; and for which I need his blessings too.

I am grateful towards my supervisor, **Dr Dinesh Kumar Vishwakarma** without whom this achievement would not have been realized. It was his valuable guidance and consistent encouragement all through my research period, which helped me to overcome the challenges that came in the way. This feat was possible only because of the unconditional support provided by his. A person with an amicable and positive temperament, he has always made himself available to clarify my doubts despite his busy schedules and I consider it as a great opportunity to do my doctoral programme under his guidance and to learn from his research experience.

My sincere regards to *Prof. Yogesh Singh, Vice-Chancellor, Delhi Technological University* for providing me with a platform for pursuing my PhD work. I express my gratitude to **Prof. Ashok De**, *DRC Chairman, Department of ECE*, **Prof. N. S. Raghava**, *Head, Department of ECE*, **Prof. S. Indu**, *Former Head, Department of ECE*, and **Prof. Kapil Sharma**, *Head, Department of IT*, for their kind support and providing necessary facilities to undertake this research. I take this opportunity to thankfully acknowledge **Prof. Narendra Kumar**, *Chief Warden, DTU*, **Prof. Rajesh Rohilla**, *HOD (T&P), DTU*, and all my teachers.

I do not have words to thank my fellow labmates that helped me whenever I needed. It is my privilege to thank **Ms Priyanka Meel**, *Asst. Professor, Department of IT, DTU*, **Dr Chhavi Dhiman**, *Asst. Professor, Department of ECE, DTU*, **Mr Gaurav Tripathi**, *Sr. Scientist CERL, BEL, INDIA*, **Ashima Yadav**, *SRF, Department*

*of IT, DTU, **Deepika**, JRF, Department of IT, DTU and **Ankit Yadav** JRF, Department of IT, DTU.*

I would like special thanks to **Mrs Sushma Vishwakarma** and **Little Diya**, for their motivational support, blessing and love. I never felt that I had missed my family on the DTU campus throughout my PhD work.

I wish to acknowledge the enjoyable company rendered by my friends, **Dr Rahul Bansal, Akhilesh Verma, Ishu Tomar, Amerendra Mishra, Prashant Rathi, Rahul Thakur , Rehan, Munindra, Ashish Kumar, Amit Gautam**, and many others who offered me their time, help and support whenever needed.

My parents are my life-coach; I feel proud of dedicating this PhD onto her lotus feet. I am grateful to my parents being an inspiration for me from childhood. They always taught me not to fear in any adverse situation and fight all odds with courage and patience.

Heartfelt and endless thanks to my sister **Dr Teena Chaudhary**, and **Dr Varun Chaudhary** (Jija Ji) and my elder brother **Dr Amit Kumar** and **Dr Babita** (Bhabhi Ji) and my **little angel Yashasvi** who always supported me at every stage of my life. They strengthen me by providing moral and emotional support. I am confident to be energized with their care, affection and brilliant advice in the rest of my life.

Tej Singh

ABSTRACT

Recently, the vision-based understanding in video sequences entices numerous real-life applications such as gaming, robotics, patients monitoring, content-based retrieval, video surveillance, and security. One of the ultimate aims of artificial intelligence society is to develop an automatic system that can be recognized and understand human behaviour and activities in video sequences accurately. Over the decade, many efforts are made to recognize the human activity in videos but still, it is a challenging task due to intra-class action similarity, occlusions, view variations and environmental conditions.

To analyse and address the issue involved in the recognition of human activity in video sequences. Initially, we have reviewed the most popular and prominent state-of-the-art solutions, compared and presented. Based on the literature survey, these solutions are categorized into handcrafted features based descriptors and automatically learned feature based on deep architectures. In this thesis work, the proposed action recognition framework is divided into handcrafted and deep learning-based architectures which are then utilized throughout this work by embedding the new algorithms for activity recognition, both in the handcrafted and automatic learned features domains.

First, a novel handcrafted feature based descriptor is presented. This method addressed the major challenges such as abrupt scene change phenomena, clutter background and viewpoints variations by presented a novel visual cognizance based multi-resolution descriptor for action recognition using key pose frames. This descriptor framework is constructed by computation of textural and spatial cues at multi-resolution in still images obtained from videos sequences. A fuzzy inference model is used to select the single key pose image from action video sequences using maximum histogram distance between stacks of frames. To represent, these key pose images the textural traits at various orientations and scales

are extracted using Gabor wavelet while shape traits are computed through a multilevel approach called Spatial Edge Distribution of Gradients (SEDGs). Finally, a hybrid model of action descriptor is developed using shape and textural evidence, which is known as Extended Multi-Resolution Features (EMRFs) model. The action classification is carried through two most famous and efficient distinctive classifiers known as SVM and k-NN. The performance of the EMRF is computed on four publically available datasets, and it shows outstanding accuracy as compared with earlier state-of-the-art approaches which show its applicability for real-time applications.

Second, two deep learning-based ConvNet architectures are presented to overcome the limitations of handcrafted solutions. These ConvNets frameworks is based on transfer learning by utilized a pre-trained deep model for features extractions to identify the human actions in video sequences. It is experimentally observed that deep pre-trained model trained on a large annotated dataset is exchangeable to action recognition task with the smaller training dataset. In the first work, a deeply coupled ConvNet for human activity recognition proposed that utilize the RGB frames at the top layer with Bi-directional Long Short Term Memory (Bi-LSTM), and at the bottom layer, CNN model is trained with a single Dynamic Motion Image (DMI). For the RGB frames, the CNN-Bi-LSTM model is trained end-to-end learning to refine the feature of the pre-trained CNN while dynamic images stream is fine-tuned with the top layers of the pre-trained model to extract temporal information in videos. The features obtained from both the data streams are fused at the decision level after the softmax layer with different late fusion techniques. The highest classification accuracies are achieved with significant margin through the proposed model on four human action datasets: SBU Interaction, MIVIA Action, MSR Action Pair, and MSR Daily Activity as compared with similar state-of-the-arts and outperforms.

In the second proposed framework, a deep bottleneck multimodal feature fusion (D-BMFF) technique is presented that utilized three different modalities RGB, RGB-D(depth) and 3D coordinates information for activity classification because it helps for better recognition and complete utilization of information available from a depth sensor video simultaneously. During the training process RGB and depth, frames are fed at regular intervals for an activity video while 3D coordinates are first converted into single RGB skeleton motion history image (RGB-SklMHI). The multimodal features obtained from bottleneck layers before the top layer are fused by using multiset discriminant correlation analysis (M-DCA), which helps for robust visual action modelling. Finally, the fused features are classified using a linear multiclass support vector machine (SVM) technique. The proposed approach is evaluated over four standard RGB-D datasets: UT-Kinect, CAD-60, Florence 3D and SBU Interaction. Our method exhibits excellent results and outperforms the state-of-the-art approaches.

Finally, this thesis work is concluded with significant findings and future research aspects in the field of human action recognition.

List of Publications

1. **Tej Singh**, D. K. Vishwakarma, "A Deeply Coupled ConvNet for Human Activity Recognition using Dynamic and RGB Images", *Neural Computing and Applications*, 2020, (Pub: Springer)- **Impact Factor: 4.774**
2. **Tej Singh**, D.K. Vishwakarma "Video Benchmarks of Human Action Datasets: A Review", *Artificial Intelligence Review*, Vol. 52, No. 2, pp. 1107–1154, 2019. (Pub: Springer Nature). **Impact Factor: 5.745**
3. D.K. Vishwakarma, **Tej Singh**, "A Visual Cognizance Based Multi-Resolution Descriptor for Human Action Recognition using Key Pose", *International Journal of Electronics and Communications*, Vol. 107, pp. 513-521, 2019, (Pub: Elsevier) **Impact Factor: 2.925**.
4. **Tej Singh**, D. K. Vishwakarma, "A deep multimodal network based on bottleneck layer features fusion for action recognition", *Concurrency and Computation: Practice and Experience*, (Pub: Wiley) (Under review).
5. **Tej Singh**, D.K. Vishwakarma, "Human Activity Recognition in Video Benchmarks: A Survey", in *Advances in Signal Processing and Communication* (Pub: Springer), 2019.
6. **Tej Singh**, D.K. Vishwakarma, "A Hybrid Framework for Action Recognition in Low-Quality Video Sequences", 2019, *arXiv:1903.04090v1 [cs.CV]*.
7. **Tej Singh, et al.**, "Deep Learning Framework for Single and Dyadic Human Activity Recognition", in *5th IEEE International Conference on Multimedia Big Data(BigMM-19)*, Singapore, September 11-13, 2019.
8. **Tej Singh**, D.K. Vishwakarma, "A Framework for Human Action Recognition using Skeleton Joints" in *IEEE International Conference on Signal Processing, VLSI and Communication Engineering*, Delhi, INDIA, March 28-30, 2019.

9. **Tej Singh** and D.K. Vishwakarma, “A Hybrid Neuro-Wavelet Based Pre-Processing Technique for Data Representation” in *IEEE International Conference on Computational Intelligence and Computing Research (IEEEICCIC17)*, Karumathampatti, INDIA December 14-16, 2017.

List of Figures

Figure 1.1: Levels of Human Activities.....	4
Figure 1.2: Block Diagram of Human Action Recognition System.....	6
Figure 2.1: A taxonomy of HAR Solutions in Video datasets.....	20
Figure 3.1: Shown Visual Cognizance Based Multi-Resolution Descriptor for Human Action Recognition using Key Pose	46
Figure 3.2: Fuzzy trapezoidal membership function for selecting the key poses frames..	47
Figure 3.3: Illustration of workflow for selecting a single image from input video of KTH dataset	50
Figure 3.4: Simulation Results of Proposed Algorithm 2	51
Figure 3.5: First row shows Region of Interest (ROI) on various activities images, Second row shows edges computed on different postures, and third row represents histogram of SEDGs at level '2'.	53
Figure 3.6: Shown the orientation features extraction vector map using Gabor Wavelet Transform	54
Figure 3.7: Procedure of proposed EMRFs framework for HAR.....	55
Figure 3.8: Example frames from Weizmann Action Dataset.....	56
Figure 3.9: Example frames of KTH Action Dataset	57
Figure 3.10: Example frames from Ballet Movement Dataset.....	58
Figure 3.11: Example frames from UCF YouTube Action Dataset.....	59
Figure 3.12: Classification result of k-NN classifier on (a) Weizmann Action (b) KTH (c) Ballet Movement (d) UCF YouTube action datasets.	62

Figure 3.13: Classification result of SVM classifier on (a) Weizmann Action (b) KTH (c) Ballet Movement (d) UCF YouTube action datasets	63
Figure 3.14: ARA Comparison of k-NN and SVM Classifier on HAR datasets	64
Figure 4.1: Schema of Deeply Coupled ConvNet for Human Activity Recognition using Dynamic and RGB Images	71
Figure 4.2: Block diagram of Inception-v3 improved Deep Architecture	72
Figure 4.3: Basic LSTM Architecture	73
Figure 4.4: The Bi-directional LSTM Architecture	75
Figure 4.5: Shown the process of calculation of parameter $\Omega\mathbf{T}$ for finite length video sequences \mathbf{T} . The bold part shows the dependency of parameter $\Omega\mathbf{T}$ on consecutive video frames $\in (\mathbf{1}, \mathbf{T})$	79
Figure 4.6: Shown the formation of Dynamic Motion Image using Approximate Rank Pooling Mechanism from Red, Green and Blue channel of each video frame.....	79
Figure 4.7: Sample RGB frames from SBU Interaction dataset.....	82
Figure 4.8: Sample RGB frames from MIVIA Action dataset.....	83
Figure 4.9: Sample RGB frames from MSR Action Pairs dataset.....	84
Figure 4.10: Sample RGB frames from MSR DAILY Activity 3D dataset.....	85
Figure 4.11: Shown the training minimum squared (MSE) loss and test loss for activity datasets: a) SBU Interaction b) Mivia Action c) MSR Action Pairs d) MSR Daily Activity.....	87
Figure 4.12: Shown results on four datasets with different data inputs: Only RGB frames, Dynamic motion image (DMI) and RGB+DMI.....	91

Figure 4.13: The confusion matrix of datasets: a) SBU Interaction b) MIVIA Action c) MSR d) MSR Daily Activity.....	93
Figure 4.14: Schema of deep multimodal network based on bottleneck layer features fusion for action Recognition.....	94
Figure 4.15: Shown the systematic block diagram for 35x35 grid module of Inception ResNet-A, 17x17 grid module of Inception ResNet-B, and 8x8 grid module of Inception ResNet-C.....	96
Figure 4.16: Shown the multiset DCA analysis on three data stream features RGB, depth and RGB-SkelMHI.....	103
Figure 4.17: Sample frames from UT Kinect dataset.....	105
Figure 4.18: Sample frames from CAD 60 Dataset.....	106
Figure 4.19: Sample frames from Florence 3D Action.....	107
Figure 4.20: Shown the confusion matrix of proposed D-BMFF technique on four datasets: a) UT Kinect, b) CAD-60, c) Florence Action 3D, and d) SBU Interaction dataset.....	108
Figure 4.21: Shown the fusion accuracies comparison obtained from different pre-trained models on RGB-D datasets.....	113

List of Tables

<i>Table 2.1: State-of-the-art Accuracy on RGB Dataset</i>	<i>39</i>
<i>Table 2.2: State-of-the-art Accuracy on RGB-D(Depth) Dataset.....</i>	<i>41</i>
<i>Table 3.1: Result comparison with the state-of-the-art on Weizmann Action Dataset</i>	<i>65</i>
<i>Table 3.2: Result comparison with the state-of-the-art on KTH Dataset.....</i>	<i>66</i>
<i>Table 3.3: Result comparison with the state-of-the-art on Ballet Dataset.....</i>	<i>66</i>
<i>Table 3.4: Result comparison with the state-of-the-art on UCF YouTube Dataset.....</i>	<i>67</i>
<i>Table 4.1: Activity wise Results of RGB Frames and DMI streams on SBU Interaction Dataset.....</i>	<i>89</i>
<i>Table 4.2: Activity wise Results of RGB Frames and DMI streams on MIVIA Action Dataset.....</i>	<i>89</i>
<i>Table 4.3: Activity wise Results of RGB Frames and DMI streams on MSR Action Pairs Dataset.....</i>	<i>89</i>
<i>Table 4.4: Activity wise Results of RGB Frames and DMI streams on MSR Daily Activity Dataset.....</i>	<i>90</i>
<i>Table 4.5: Wilcoxon Rank Sum Test results on four human activity datasets.....</i>	<i>90</i>
<i>Table 4.6: Accuracy (%) comparison of different fusion techniques on human activity datasets.....</i>	<i>91</i>
<i>Table 4.7: State-of-the-art comparison of SBU interaction Dataset</i>	<i>94</i>
<i>Table 4.8: State-of-the-art comparison of MIVIA Action Dataset.....</i>	<i>95</i>
<i>Table 4.9: State-of-the-art comparison of MSR Action Pairs Dataset.....</i>	<i>95</i>
<i>Table 4.10: State-of-the-art comparison of MSR Daily Activity Dataset.....</i>	<i>96</i>

<i>Table 4.11: Comparison of accuracies of Inception-Resnet-v2 architecture with similar state-of-the-art architectures</i>	<i>99</i>
<i>Table 4.12: Comparison of multimodal features fusion score using Inception-Resnet-v2 on UT Kinect Dataset.....</i>	<i>110</i>
<i>Table 4.13: Comparison of multimodal features fusion score using Inception-Resnet-v2 on CAD 60 Dataset.....</i>	<i>110</i>
<i>Table 4.14: Comparison of multimodal features fusion score using Inception-Resnet-v2 on Florence 3D Action Dataset.....</i>	<i>111</i>
<i>Table 4.15: Comparison of multimodal features fusion score using Inception-Resnet-v2 on SBU Interaction Dataset</i>	<i>112</i>
<i>Table 4.16: Results comparison of Pre-trained architectures on UT Kinect Dataset</i>	<i>113</i>
<i>Table 4.17: Results comparison of Pre-trained Architectures on CAD-60 Dataset</i>	<i>114</i>
<i>Table 4.18: Results comparison of Pre-trained Architectures on Florence 3D Dataset</i>	<i>114</i>
<i>Table 4.19: Results comparison of Pre-trained Architectures on SBU Interaction Dataset</i>	<i>114</i>

Table of Contents

DECLARATION	ii
CERTIFICATE.....	iii
ACKNOWLEDGMENT.....	iv
ABSTRACT	vi
List of Publications.....	ix
List of Figures.....	xi
List of Tables.....	xiv
Chapter 1 Introduction of Human Activity Recognition.....	1
1.1 Background.....	1
1.2 What is Action?.....	3
1.2.1 Modality.....	4
1.3 Human Action Recognition.....	5
1.4 Challenges in HAR.....	8
1.4.1 Background and Environment Conditions.....	9
1.4.2 Intra and Inter-class Variations.....	9
1.4.3 Occlusion	10
1.4.4 View-Variations	10
1.4.5 Lack of Labelled Data	10
1.5 Applications.....	11
1.5.1 Interactive Applications and Environments.....	11
1.5.2 Behaviour Biometric	11
1.5.3 Content-based video retrieval.....	12

1.5.4 Animation and Synthesis	12
1.5.5 Video surveillance and security.....	12
1.6 Problem Statement	13
1.7 Major Contributions of Thesis.....	14
1.7.1 Theoretical Formulation.....	14
1.7.2 Experimental Validation	15
1.8 Motivations	16
1.9 Significance of Human Action Recognition	17
1.10 Thesis overview	17
Chapter 2 Literature Review	19
2.1 Introduction	19
2.1.1 Handcrafted features based Solutions	21
2.1.2 Global Features Extraction based Representation.....	21
2.1.3 Local Features Extraction based Representation.....	23
2.1.4 Still Image Based Action Recognition	30
2.1.5 Hand Crafted Feature Descriptors based on Skeleton Sequences	31
2.1.6 Spatio-Temporal Features based Architectures	33
2.1.7 Multiple Stream Network-based Architectures	35
2.1.8 Deep Generative Network-based Architectures	36
2.1.9 Temporal Coherency Network-based Architectures	37
2.2 State-of-the-art Accuracy on RGB and RGB-D Datasets	38
2.3 Gaps Identified in the Present Study	42
2.4 0.Research Objectives	43
Chapter 3 Human Activity Recognition Using Handcrafted Features	44

3.1 Introduction	44
3.2 A Visual Cognizance Based Multi-Resolution Descriptor for Human Action Recognition using Key Pose	45
3.2.1 Selection of Single Key Pose using Fuzzy Logic	45
3.3 Extended Multi-Resolution Features (EMRFs)	49
3.3.1 SEDGs Feature Map.....	49
3.3.2 Orientation Feature Map	52
3.4 EMRFs Representation.....	53
3.4.1 Performance Evaluation on Human Action Datasets	56
3.4.2 Weizmann Action Dataset.....	57
3.4.3 KTH Action Dataset.....	57
3.4.4 The Ballet Dataset	58
3.4.5 UCF YouTube Action	59
3.5 Experimental details and Results Analysis	59
3.6 Comparison of EMRFs with State-of-the-Art Approaches	64
3.7 Significant Outcomes	67
Chapter 4 Learned features based Action Recognition	69
4.1 Introduction	69
4.2 A Deeply Coupled ConvNet for Human Activity Recognition using Dynamic and RGB Images.....	70
4.2.1 Features Extraction with Pre-Trained Inception-v3 Architecture.....	71
4.2.2 The Bi-Directional LSTM (Bi-LSTM)	73
4.2.3 Dynamic motion image (DMI) from video sequences	76
4.2.4 Late Fusion	80

4.2.5 Implementation Details	81
4.2.6 Model Parameter Description and Training Settings	86
4.2.7 Results Analysis and Comparisons	88
4.2.8 The Mann-Whitney U Test (Wilcoxon Rank Sum Test).....	88
4.2.9 Results comparison with State-of-the-art.....	91
4.3 A deep multimodal network based on bottleneck layer features fusion for action recognition.....	97
4.3.1 Deep fusion framework for Human Activity Recognition.....	98
4.3.2 Implementation Details	104
4.3.3 Result Analysis.....	106
4.3.4 State-of-the-art Comparison.....	109
4.4 Significant Outcomes	115
Chapter 5 Conclusion and Future Scope	117
5.1 Conclusions.....	117
5.2 Future Prospective	120
5.3 Future Applications	121
References.....	123
Author Biography	143

Chapter 1

Introduction of Human Activity Recognition

Human activity recognition in videos became an imperative choice of researchers in computer vision because of its wide range of real-life applications. It includes e-health, patients monitoring, assistive daily living activities, video surveillance, security and behavior analysis, sports analysis and many more. This chapter introduced the background of the human activity recognition system, basic terminology, fundamental architecture, various challenging presents in video analysis, and numerous applications of human action recognition in day to day life. Furthermore, research problem statements, significant contribution, motivations for research, significance of the study, and thesis organization are discussed.

1.1 Background

Human activity recognition in the video sequence is the most popular and ever-increasing area of computer vision research due to its plethora of applications in daily life. It includes safety, surveillance, healthcare, robotics, animations, sports analysis, content-based video summarization, and behavioural analysis, smart homes and many more. One of the ultimate aims of artificial intelligence society is to develop an automatic system that can recognize and understand human behaviour and activities accurately. So that it can serve the society in a better way for example, a robot assistant can be capable enough of assisting a patient under observation at home and analysing the right way of exercise and preventing the patient from future injuries. Therefore, such an intelligent system will be very helpful

for us as it saves time to visit the doctor, reducing the medical cost and provides continuous remote monitoring of the patient. In the past two decades, many hand-crafted and automatically learned feature-based approaches developed for human action recognition in the videos. Earlier human activity recognition approaches are based on handcrafted features mainly focused on simple atomic actions [1] [2] [3]. Later on, convolutional neural networks(CNNs) based deep models for video activity analysis were proposed that can automatically learn the features and classify from raw video only [4] [5] [6] [7].

The handcrafted feature extraction approaches for activity recognition is based on spatial background subtraction, optical flow, dense trajectories, and human pose variations [8] [9] [10] [11] [12]. After the tremendous progress of deep learning architectures on human pose estimation [13], object detection [14], segmentation [15], speech analysis [16], object tracking [17] and super-resolution [18]. The deep learning model also plays a central role in visual recognition tasks. Unlike handcrafted solutions, deep learning-based approaches provide a new way to extract the features from images automatically. It is observed that handcrafted features solutions showed promising results but relied more on features descriptors for action classification. These solutions required more labour and subject knowledge expertise. On the other hand, the deep learning-based approaches are dominated because of automatic features extraction from raw videos and provides better recognition rate.

Still, human activity recognition is a challenging problem in machine learning and many key difficulties remain unresolved such as intra-class variation, illumination changes, occlusion, actions similarities, viewpoint variations, change in scale, appearance, age, frame resolutions, and lightening conditions [19] [20] [21] [22]. With the invention of advanced Kinect depth sensor various deep learning methods based on single modality (RGB, depth(D), and skeleton coordinates) and their various combinations are introduced [23] [24] [25] [26] [27]. However,

very few approaches are based on the combination of RGB, depth and 3D-skeleton coordinates for activity recognition [28].

In general, a video consists of visual multimedia information in the form of sequences of images (frame per second). Unlike the feature representation in an image, the human action modelling in video sequences is based on Spatio-temporal features representation. The spatiotemporal models extracted the spatial appearance features present in video frames and existing pose variations. The extraction of Spatio-temporal features is needed to recognize the action in video sequences. The main objective of this thesis work is to automatically detect and analyse human action or activities from the data acquired from sensors, e.g. video camera, depth sensors and other modalities.

1.2 What is Action?

According to Oxford dictionary, “the fact or process of doing something, typically to achieve an aim” is called **action** and similarly, an **activity** is defined as “a thing that a person or group does or has done”. There are numerous definitions of action given by various authors in their works. However, most suitable stated by Herath et al. [1], “*Action is the most elementary human-surrounding interaction with a meaning*”. Therefore, human activities can be classified based on interaction with the surrounding into four major categories as follows:

- *Gestures* are defined as basic or atomic movements of body parts, for example, hand waving, moving the head up and down, etc.
- *Actions* are defined as a body activity of one person or actors such as running, walking, and jogging, etc. An action can be considered as combinations of atomic gestures.
- *Interactions* are defined as two or more peoples involved in one to one communications. It may be both types of human-human interaction (HHI) and

human-objects interaction (HOI). Examples of human-human interactions are handshaking, hugging, exchanging objects fighting with each other, etc. while Human objects interactions (HOI) may see as the person making tea, answering the phone calls etc.

- *Group Activities* are defined as multiple persons or groups involved in common objectives such as a group meeting, two groups fighting with each other, etc.

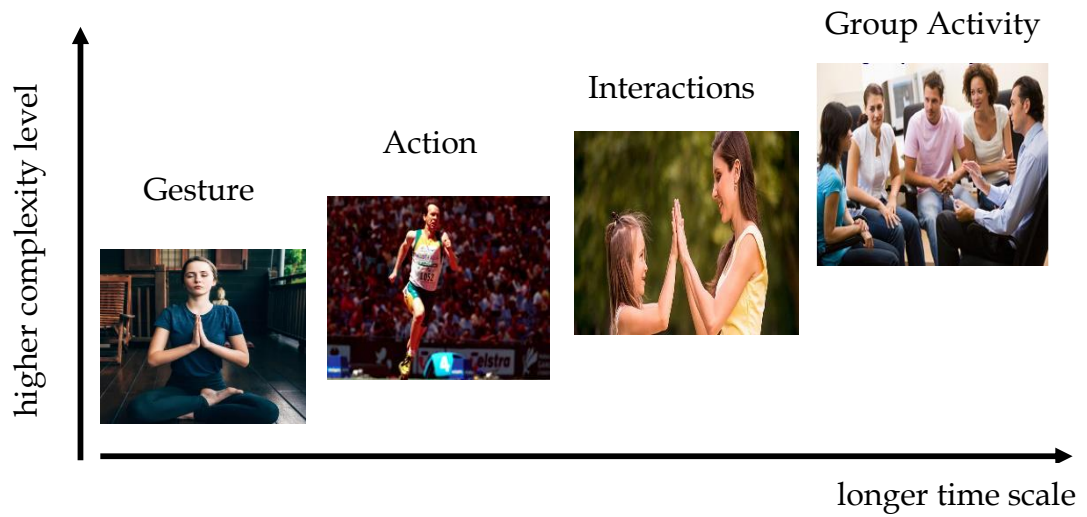


Figure 1.1: Levels of Human Activities

It can be observed from Figure 1.1 that the complexity of action representation is increased from gesture recognition to group activity recognition.

1.2.1 Modality

It is defined as a method to record and collect information in a video dataset. A multimodal video dataset is recorded with two or more sensors to captures the human activity in the videos. The three different types of modalities that collect the information from depth sensors are as follows:

- *RGB frames:* It is available in various resolutions and formats. The format determines the type of colour image data, whether it is encoded as RGB or

Grey. High-resolution images have high data per frames as compared to low-resolution images.

- *Depth frames*: It is a measure of distance, in millimetre, to the nearest object at that particular spatial coordinates in the depth sensor's (Kinect) field of view. The depth images can be captured in three different spatial resolutions: 640×480, 320×240, and 80×60 according to specified image format. It can be utilized to track a person's motion and background segmentation tasks.
- *3D Skeleton coordinates*: It consists of the 3D position of data for human skeletons which are visible in the depth sensors. In tracking mode, the position of a skeleton and each of the skeleton joints is stored in three-dimensional coordinates in meters.

This thesis work mainly focused on recognition of action and interaction activities from video sequences by utilizing all three modalities either individually or combinedly.

1.3 Human Action Recognition

Typical human activity recognition can be broadly categorized into three different representation levels, as depicted in Figure 1.2. It consists of low-level fundamental technology to extract the information from sensors, mid-level various activity recognition and associated high-level human activity recognition applications.

In a traditional HAR system, low-level generally represented the core fundamental techniques that are necessary for human activity recognition from video sequences. It involves various basic steps, such as pre-processing of input frames, features extraction, and action classification. On the other hand, deep learned features based methods extracted the features from pixels' basis and classified the activity automatically. The basic steps utilized by the core technology from data acquisition to representation and classification of human activity are as follows:

- *Pre-processing*: It is used to enhance the quality of input video sequences for robust features extraction. It includes various techniques such as background segmentation, silhouettes extractions, histogram equalization, optical flow estimation etc. Earlier human action recognition methods focused on the processing of extraction of human silhouettes to represent human motion. It includes background removal, frame normalization, and vector quantization in a constrained environment.

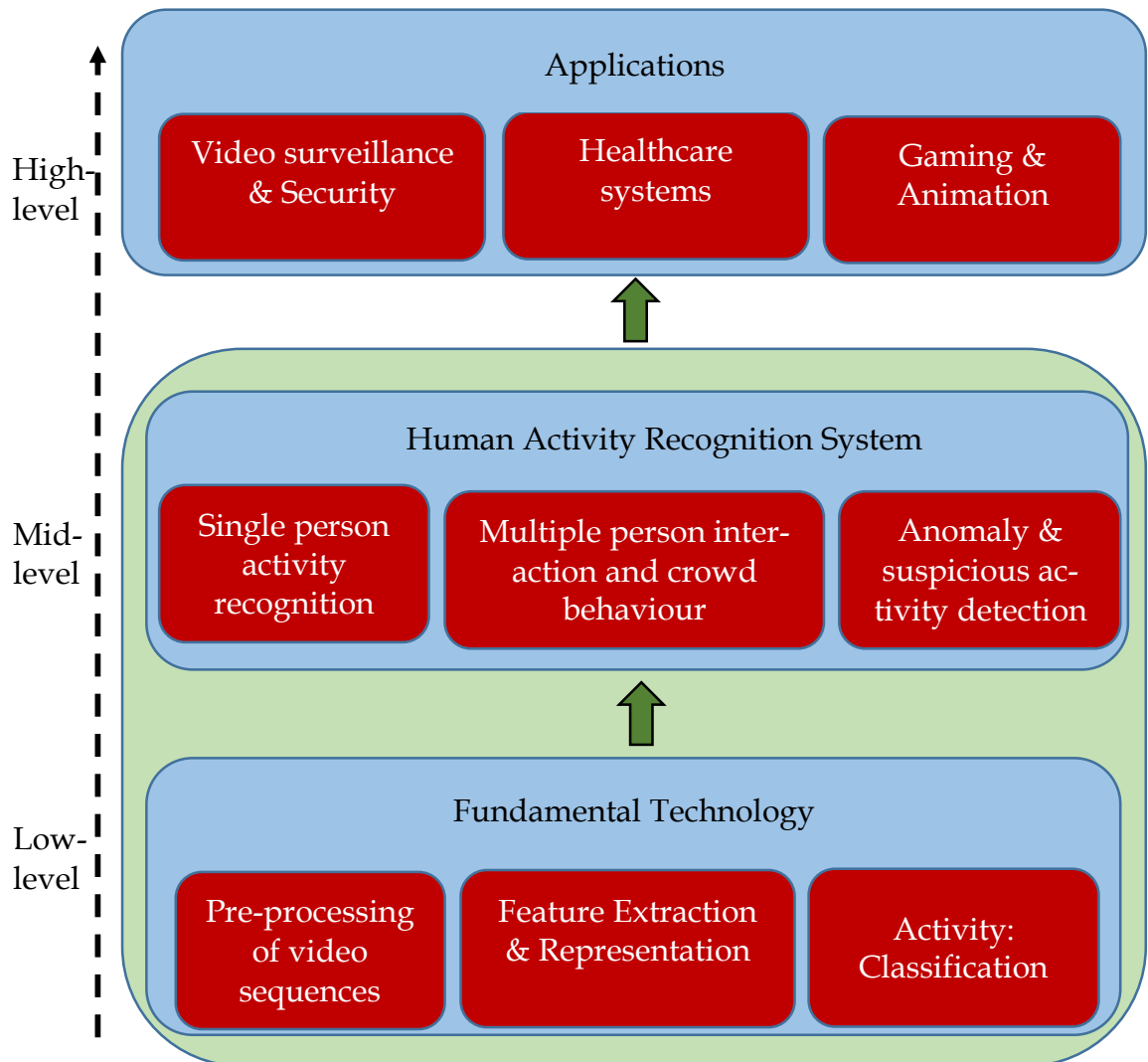


Figure 1.2: Block Diagram of Human Action Recognition System

The main disadvantages of these pre-processing techniques are that these are less applicable to real-time applications and not so efficient for unconstrained environments such as complex and low illumination conditions. With the invention of advanced depth sensors that provide different modality including RGB frames, depth and 3D skeleton coordinates. The background segmentation is easier in depth frames as compared with RGB frames for object recognition in the presence of clutter background or illumination variation. The 3D skeleton joints represent the human posture movements by tracking the human-specific parts such as feet, hands, legs, arms, and torso. The pre-processing of these 3D features are robust and invariant to illumination and background conditions that can be helpful for real-time action recognition.

- *Feature extraction and Representation:* The raw video sequences acquired from sensors contains redundant information. The feature extraction process removed such redundant information from raw videos and unveiled the hidden Spatio-temporal relationship to recognized human activity. Further, the feature extraction techniques eliminate the noise that occurred during the recording of data from sensors. It will reduce the memory requirement and save precious time for classification. There are various well known and established hand-crafted feature extraction methods such as MHI, MEI, STIP, SIFT, Optical flow, BoWs, HOG, HOF, dense trajectories etc. to extract the robust features for activity representation. Unlike the traditional features, deep learned models extracted features automatically from raw pixels of video sequences. Transfer learning is well known approach to extract the features from input sensors. Other features extraction techniques used for sequential information extraction are Recurrent Neural Network (RNN), LSTM (long short term memory). The automatic learned feature models are dominated over hand-crafted features due to their more robust architecture and generalization to real-life problems.

- *Activity Classification:* It is the final step in the activity recognition system. The classification accuracy directly depends on the features extraction process from input data. Traditional hand-crafted methods utilized the machine learning techniques for label the activity classes. These techniques are linear multiclass SVM, HMM, K-NN, Random forest, Bootstrap and k-means etc. Automatic learned deep methods used the Softmax classifier after the fully connected layers to classify the action class. Recently, multi-streams fusion networks with different fusion techniques such as early and late fusion are used for final prediction.

The mid-level of the HAR system consists of the single-person activity, multiple person activities, crowd behaviour, anomaly detection and suspicious activities. This level extended the low-level human action recognition to be more specific into realistic problems such as assistive daily living applications and human behaviour understanding with the surrounding. At last, high level HAR system discussed the necessity of human activity recognition and various potential applications based on video analysis. These HAR applications play a significant role in day to day life such as video surveillance, security, healthcare, sports and many more.

In the following sub-section, the various challenges involved in human activity recognition in video sequences are discussed in details.

1.4 Challenges in HAR

The recorded videos in the datasets are having limitations in at least one of aspects such as similarity of actions, cluttered background, viewpoints variations, illuminations variations, and occlusions. Initially, human action datasets were less challenging compared to current dataset due to less number of action classes, subjects,

and reasonably recorded in controlled environmental conditions. The performance measure of an algorithm depends on these crucial factors present in the datasets. The RGB datasets is more challenges like view changes, intra-class variations, cluttered background, partial occlusions, and camera movements than RGB-D(depth) datasets.

1.4.1 Background and Environment Conditions

The natural environment contains various objects such as trees, waves, rain, and water and these factors affect the recognition activity in videos. The performance of feature descriptors directly influenced by background subtraction techniques or foreground detection. The KTH Action dataset [29] is more challenging due to changing background compared to Weizmann dataset [30]. Recognizing human activity in videos is a crucial task in the presence of moving object or background. The background in videos may be of different types such as slow/fast, dynamic, static, occluded, airy, rainy, and densely populated.

1.4.2 Intra and Inter-class Variations

It can be noticed that different persons performed same actions in a different manner. For considering the ‘running’ action, a person can run slow, fast or sometimes jumps and then run. It means an action class may contain different styles performed by the human motion. Furthermore, the execution time of action vary person to person pose variations. All such factors lead to interclass pose and appearance variations [31] [32]. The similarity between the actions classes in videos provides a fundamental challenge to the researcher. Many actions seem to be similar in videos such as jogging, running, walking, etc. These similarities provide the challenges for automatic recognition system which lead to misclassification.

1.4.3 Occlusion

Occlusion is a thing where another object hides the object of interest. It is a challenging task to recognize human activities from occluded videos. The occlusion is a major challenge in the field of computer vision such as human pose estimation, object tracking, video surveillance, 3D foreground reconstruction, and traffic monitoring applications. It occurs due to the relative motion of static and dynamic occluding objects. Occlusion is two types in human pose estimation self-occlusion and occlusion by another object. Self-occlusion is occurring when body parts occlude each other due to different viewpoints while other occlusions occur when an object obstructs the view [33] [34].

1.4.4 View-Variations

The viewpoint of any activity recorded inside the video dataset is a key attribute in the human activity recognition system. The videos recorded with multiple views have more robust information than a single view and independent of the captured view angle. However, multi-views increase the complexity of the HAR system [3] [35].

1.4.5 Lack of Labelled Data

It is observed that most of the HAR approach shown impressive performance on small human activity datasets. It is a challenging task to generalized these solutions on large scales for real-time applications. The deep network based architectures provide a promising performance on large scale datasets. But these deep models required a large amount of labelled training data for training the model [36] [37]. Although, few action datasets such as YouTube-8M [38] and Sports-1M [32] consist of millions of action videos. But annotations of videos are created by retrieval techniques that may not be so accurate. The training on such datasets is

challenging due to insufficient labelled training data. Therefore, the performance of action descriptors is affected due to inaccurate labelling.

1.5 Applications

HAR using video sequences spreads dimensions in a vast area of research due to its practical applications. These are broadly categorized as:

1.5.1 Interactive Applications and Environments

Human-computer interaction is a challenging task in the design of the human-computer interface. The visual features are the main components of non-verbal communication. Human activity recognition such as gestures, actions, and interactions can be utilized as a useful tool for this interaction. A HAR model helps to a better understanding between the human-computer interface such as robotics [39], smart homes [40]. Although these solutions are not so developed to interact perfectly; therefore, it attracts a lot of research attention.

1.5.2 Behaviour Biometric

Initially, traditional biometric algorithms are based on physical attribute cues such as iris, fingerprints or face to recognizing the human identity. It is noted that these solutions need the physical involvement of the person to be identified. Due to such limitations, 'Behavioral Biometric' attracts much attention to recognize person identity because it is observed that the behaviour of a person is an essential cue as equal to other physical cues. Further, it has an advantage because of no interference of person to recognizing the personal activities. In this context, to identify a person behaviour need more observation time. Therefore, human activity recognition in video sequences plays a significant role in behavioural biometric systems such as Gait Analysis [41].

1.5.3 Content-based video retrieval

Today, video plays a core role in the day to day lifestyle to sharing information on social media platform (Facebook, Twitter, Instagram) or online entertainment websites (YouTube, Netflix, hot star). The video summarization and indexing are gaining more popularity similarly the content-based image retrieval(CBIR) because of utilization in potential commercial field application such as sports analytics [42] [43] [44].

1.5.4 Animation and Synthesis

The human motion analysis is useful for gaming industry due to variation of poses, view variations, and motion patterns. The movie industry highly depends on good quality animation applications that depend on fusing both the realistic human and human motion. This is fast-growing area many applications are developed for military applications, fire-fighters and other national disaster management team for rescue purposes in adverse situations. With the help of advanced algorithms and high computing devices, a simulated environment is created for training these soldiers for hazardous conditions. Due to the availability of high-quality videos, inexpensive hardware and continuous monitoring, it is possible to track the desired target for object detection. Therefore, a HAR system helps to improve the qualities of the existing challenging problem [45].

1.5.5 Video surveillance and security

The security and surveillance system in our houses and smart city directly depend on videos, e.g. CCTV cameras installed on a specific location. Traditional security networks are highly depending on the awareness of security persons and the camera's area of view. Recently, with the availability of inexpensive high resolution capturing devices and connectivity of internet stretched the vision-based tasks and

eliminated the dependency of operators. The advance security networks are seeking automatic recognition of suspicious or abnormal activities. That is why automatic human activity recognition in video sequences attracted the researcher's attention to computer vision-based applications more rapidly [46] [47].

1.6 Problem Statement

Based upon the challenges mentioned above such as intra-class similarity, view variation, scales, varying illumination conditions, clutter background, and various types occlusions, need to developed a HAR system that can overcome such limitations exist in video sequences. To handle such issues an effective, robust models based on handcrafted features descriptor and learned features based architectures are presented in this thesis work. Practically to handle the problem of action recognition in videos, we have formulated the following statements given below as:

- To design and develop a robust activity recognition model that can automate analysis or interpretation of ongoing events and their context from video data.
- To design an algorithm which can handle issues such as low illumination, de-noising, background and enhance the video quality so that representation is more efficient to recognize human activities.
- To design and develop a deep learning-based model for action classification that is invariant to scale, viewpoint, occlusions, illuminations changes and environmental conditions.
- To design and develop a multimodal features fusion approach based on deep convolution neural network architecture that can automatically learn the Spatio-temporal features for efficient human action modelling in video sequences.
- To validate the developed algorithm and experiment can be conducted on standard human activity RGB-D datasets and know the effectiveness of the

novel algorithm through a comparative study and implementation to be conducted.

1.7 Major Contributions of Thesis

This section explained the major contribution of this thesis work. This work is also presented the theoretical formulation and experimental validation basis for improvement of HAR solutions in the following sub-sections.

1.7.1 Theoretical Formulation

The theoretical contribution of this work as follows:

- The problem of the frame redundancy is identified in videos sequences, and an appropriate model is chosen which can adequately represent the video with a small set of discriminative key pose frames.
- The issue of efficient shape features extraction and noise removal from a background in video frames is identified and solved with the help standard edges detector with a threshold mechanism.
- To handle the issue of view variation and scales variations, an appropriate orientation model is chosen at various orientations and scales.
- The problem of low recognition accuracy in the HAR model under various challenging conditions is detected.
- The issue of overfitting the trained model due to small samples datasets is addressed by utilizing the concept of transfer learning and dropout mechanism for extracting the Spatio-temporal features for action representations.
- The issue due to limitations of joints and coordinates of upper or lower body parts used to represent activity is studied and dealt with.
- To handle the issue of skeleton data is degraded due to noise and occlusions present in RGB-D images has been identified.

- The performance of various fusion techniques under different constraints for effective features extractions has been observed.

1.7.2 Experimental Validation

The proposed frameworks are experimentally evaluated over publically available standard RGB-D(depth) human activity datasets. These datasets are included single person, multiple people, human-human interaction and human-object interaction activities videos recorded in various challenging environmental conditions.

- An Extended Multi-Resolution Features (EMRFs) model is developed by concatenation of shape and textural evidence and the performance of the EMRFs is measured in terms of accuracy on standard datasets.
- Design a fuzzy-model-based approach used to select single key pose action images from input video sequences.
- Utilized Gabor wavelet transforms for extractions of the textural features at various orientations and scales.
- Experiment studies of the proposed method on the set of the reference dataset.
- A hybrid two-stream deep ConvNets are presented that utilized two different spatial and temporal data streams to recognize human action.
- Developed a hybrid two-stream deep ConvNets that utilized two different spatial and temporal data streams to recognize human action.
- Design a deep architecture based on transfer learning model for features extraction and Bi-LSTM architecture for sequential data modelling.
- Studied and observed the effect of various late fusion techniques at top layers.
- Design the bottleneck features extraction model by fine-tuning the latest pre-trained architectures for action modelling.

- Development of multimodal data stream fused with multiset DCA technique.
- Experiment studies of the proposed method on a set of the reference dataset.

1.8 Motivations

Over the last decades, understanding human activities in video sequences is tied to complementary research including human dynamics, semantic segmentation, objection recognition, and domain adaptations. Today, human action recognition can automatically learn from thousands of videos and applied to all daily life applications.

Recently, it is observing a rapid increase of video contents on social media platforms such as YouTube, Facebook and Twitter. The availability of inexpensive, high-quality camera devices and high internet speed in smart mobile phones, a huge amount of videos uploaded every year on these social media platforms. Due to the enormous amount of data, there is a need for a system which can accurately analyse these videos and provide necessary solutions and suggestions. Human action recognition is a key ingredient of such systems.

We can observe that robotics is an interdisciplinary field of science and engineering. It deals with the development of machines which can replace humans. Robots is a multipurpose machine that can be used in adverse situations such as bomb detection and de-activation or where humans could not handle. Human action recognition system plays a significant role in many robotics applications. For example, autonomous vehicles that are a specific type of robot which can control the road situation and reduce the accidental loss on roads. Autonomous driving requires an accurate human pedestrian detection and prediction of body, and it can avoid potentially dangerous situations.

With the broader range of applications from robotics to human-computer interaction and video surveillance motivated us to recognize the human actions in video sequences.

1.9 Significance of Human Action Recognition

Vision-based human action recognition is a lively area of research in the field of computer vision and machine learning. The main objective of human action recognition is to automatically detect and analyse of human activities from the data acquired from sensors, e.g. video sequences, depth sensors and other modalities. It has countless applications such as security and surveillance, assistive healthcare, human-computer interaction, robotics, user-interface design, video browsing, sports analysis, human object tracking, robotics, and prevention of terrorist activities etc. In the present societal situation, suspicious activities, road accidents, terrorist attacks, riots, and stampede are progressively increased daily. Due to a large amount of information extracted from video sequences, a HAR model can be utilized as an effective tool to combat such security issues.

1.10 Thesis overview

This thesis is organized into five chapters. The brief outlines are given below:

- **Chapter 1:** This chapter introduced the background of the human activity recognition system, basic terminology, fundamental architectures, various challenges in video analysis, applications of action recognition. Furthermore, research problem statements, significant contribution, motivations, and significance of the study are discussed.
- **Chapter 2:** This chapter explained the merits and demerits of existing state-of-the-art methods. We have reviewed the traditional hand-crafted features as well as automatic learned features descriptors for the representation of

human activity in video sequences. It helps us to discover the research gaps in existing solutions in the relevant area. We also provide a comparison for publically available human activity datasets. Further, the research objectives are formulated based on these research gaps later on which are addressed in this thesis.

- **Chapter 3:** This chapter presents a hand-crafted features descriptor for human action recognition using key pose. The proposed framework is constructed by computation of textural and spatial cues at multi-resolution in still images obtained from videos sequences, which is known as Extended Multi-Resolution Features (EMRFs) model. The effectiveness of the proposed approach is explained and validated through experiments on standard datasets and state-of-the-art comparison of obtained results.
- **Chapter 4:** This chapter introduced the two automatic learned deep frameworks for human activity recognition in RGB and RGB-D(depth) videos. The first deeply coupled ConvNet model based on transfer learning that utilized RGB only frames and dynamic images for the representation of complex actions in videos. On the other hand, our second approach utilized and fused three different modalities RGB, RGB-D(depth) and 3D coordinate information for activity classification for better action recognition and complete utilization of information available from a depth sensor video simultaneously. Further, the classification results of both deep learning approaches are validated on standard depth action datasets and compared with existing state-of-the-art methods.
- **Chapter 5:** This chapter provides a summary of proposed works, significant finding, contributions and limitations. In this chapter, we also suggest some future directions, short-term and long-term perspectives for human activity recognition in videos.

Chapter 2

Literature Review

This chapter explained the merits and demerits of existing state-of-the-art methods. We have reviewed the traditional hand-crafted features as well as automatic learned features descriptors for the representation of human activity in video sequences. It helps us to discover the research gaps in existing solutions in the relevant area. We also provide a comparison for publically available human activity datasets. Further, the research objectives are formulated based on these research gaps later on which are addressed in this thesis.

2.1 Introduction

In the last decade, many handcrafted and automatically learned feature-based approaches developed for human action recognition in the videos. Earlier human activity recognition approaches are based on handcrafted features mainly focused on simple atomic actions that seem to be somewhat less useful to practical applications [2] [9] [12] [23]. The main drawback of these approaches is data pre-processing and difficult to generalize in real life despite gaining a high accuracy model. Later on, after the success of convolutional neural networks(CNNs) on text and image classification, various Spatio-Temporal approaches for video activity analysis were proposed that can automatically learn the features and classify from raw RGB video only [4] [5][21]. However, such approaches could not achieve higher accuracy due to data dependency for training the CNN models and the lack of hardware resources [26] [27] [48] [49]. Therefore, it is required that developed

solutions should overcome the challenges present in video datasets such as cluttered background, view-variation, occlusion, intra-class similarity and application scenarios.

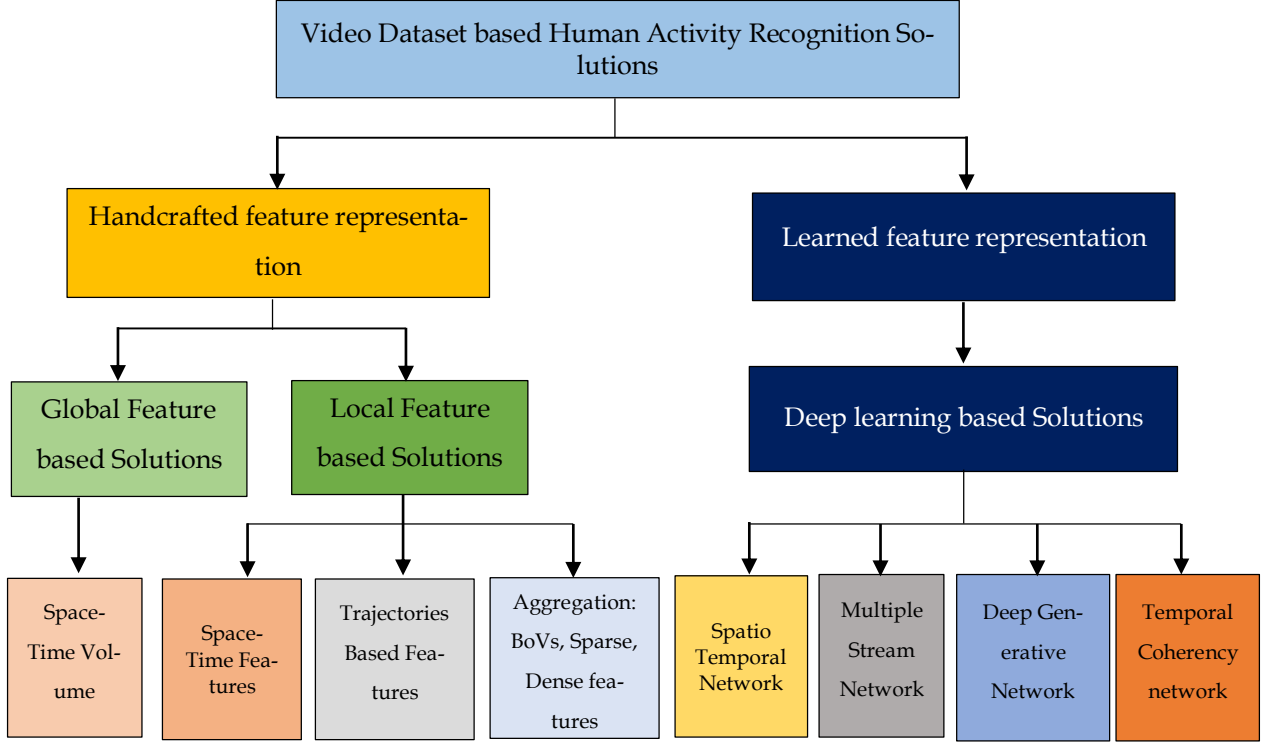


Figure 2.1: A taxonomy of HAR Solutions in Video datasets

These solutions can broadly categorised into two parts: Handcrafted feature descriptor and Automatic learned features based solutions as shown in the Fig.2.1. Handcrafted features solutions based on spatiotemporal volume dominated the research from 2001 to 2010. The main disadvantages of these descriptors solutions are average performance for the complex dataset and less generalization of the algorithm to realistic scenarios. Later on, deep learning-based solutions outperform handcrafted solutions due to robust feature extraction and classification in videos. With the advancement of high computational power and increasing size of the video dataset, deep learning-based solution is beneficial for real-life applications.

2.1.1 Handcrafted features based Solutions

The very first step towards action recognition from a video sequence is introduced by Hogg [50]. A WALKER model based on 3D structural hierarchical modelling is proposed to interpret human actions. A similar, approach based on the connected cylindrical shape to represent the limb connection for pedestrian recognition is introduced by Rohr [51]. However, building the perfect 3D model from videos is a cumbersome and costly task. Therefore, many approaches avoided 3D modelling instead developed the handcrafted features extraction techniques. The handcrafted feature-based solutions extracted the global and local features such as edge, shape, and motions from the human body. Further, these representations can be categorised into two categories depending on features extraction from the input video as follows:

2.1.2 Global Features Extraction based Representation

Global solutions are based on the features extraction of body shape and motion to represents human action.

Bobick and Davis [30] extracted the motion feature from videos sequences in the form of Motion History Images (MHI) and Motion Energy Images (MEI) temporal template to recognized human action in static background conditions. They have targeted the view of specific human motion activities and motion is considered overtime. The binary cumulative motion image for the video image sequences, MEI is defined as:

$$E_{\tau}(x, y, t) = \bigcup_{i=0}^{\tau-1} D(x, y, t - i) \quad (2.1)$$

where, E_{τ} and $D(x, y, t - i)$ represented the formed MEI and binary image sequence denoting the region of motion at a time τ respectively. The MHI template

represented the temporal history of motion denoting the pixel intensity at that location and defined as:

$$H_{\tau}(x, y, t) = \begin{cases} \tau & \text{if } D(x, y, t) = 1 \\ \max(0, H_{\tau}(x, y, t - 1) - 1) & \text{otherwise} \end{cases} \quad (2.2)$$

Both MEI and MHI template gives useful information about the moving object and removes the cluttered background problem in video sequences.

Tian et al. [52] utilized the gradient of MHI template to improvised activity recognition in a cluttered background. They detected the interest point with the help of 2D Harris corner detector [53] at the high-intensity point in MHI template. Further, the Spatio-temporal features are represented by HOG and actions are classified with the GMM model. Blank et al. [54] proposed the Space-Time 3D shape model of MEI template using binary silhouettes which outperform over previous approaches of human action recognition, detection and clustering. Their approach does not require video alignment and can be applicable for realistic scenarios. A similar approach is adopted in work proposed by Weinland et al. [55] using spatiotemporal volumes to recognized view-invariant human activities in videos. They introduced a Motion History-Volume(MHV) to represent human action view-free in multi calibrated, segmented background videos. Further, to reduce the dimensions, PCA and LDA algorithm is used and the Fourier transform is utilized to discarding the phase to recognised the primitive action classes.

Shechtman and Irani [56] proposed a behaviour based similarity matrix template to measure the similarity between human actions. They extended the 2D image correlation to 3D space-time volume to correlate the dynamic behaviour and actions. Rodriguez et al. [57] introduced a maximum average correlation height (MACH) filter template-based approach to recognize the action in videos. Their model is capable of addressing the problem of intra-class variations at a minimum computational cost. Yilmaz and Shah [58] proposed action sketch to analysed the

Spatio-temporal variations by using differential geometrical surface properties. A space-time volume (STV) is created by stacking the consecutive contours along time axis using the graph-theoretical method. Action sketch extracted the features from the surface of STV to recognised the actions and invariant to viewpoint variations.

It is noted that the global features based action representation solutions were most popular between the period of 2001 to 2007. However, at present local features and deep learning based approaches dominated the research for action recognition. The reason for this shifted focus is apparent that global features solutions are less sensitive to challenges that exist in a video such as view variations or occlusions. Furthermore, the binary silhouettes based features extraction technique is not so useful to captures the fine details in the video sequences. Therefore, the gear is changed from global to local features extraction based action representation.

2.1.3 Local Features Extraction based Representation

These techniques are based on the extraction of local features from the body shape. These techniques outperform over global feature-based techniques due to robust feature extraction and invariant to changes. These features descriptors used both cuboids as well as trajectories for action recognition and classified as follows:

2.1.3.1 Interest Point Detection Approaches

In the image pre-processing techniques, corners are considered the most important features points because of invariant to rotation, translation, and illumination changes. The two most famous techniques used to detect the interest points are Harris corner detector [53] and Hessian detector in computer vision applications. Harris corner detector used the differential gradient to detect the corners and edges while the Hessian detector detects the corner features with second-order

derivatives function. In order to recognize the action based on local interest points, Laptev et al. [59] proposed a Space-Time Interest Points(STIPs) approach by extended the 2D Harris detector to a 3D corner detector. STIP features based representation shown excellent results to pose estimation in the presence of occluded background and view variations conditions. However, their approach is sensitive to motion of camera, i.e. camera jitters. In other of extension of 2D Hessian detector to 3D, Willems et al. [60] proposed an approach to localized action using second derivatives of the corner detector. Dollar et al. [61] introduced the spatial filtering approaches to recognize and characterize the behaviour from the video sequences. They suggested that direct computation of 3D counterpart to commonly used 2D interest point are not a good idea for identifying the Spatio-temporal features.

It is a challenging task to extract the exact and informative features in the presence of camera motion and background clutter in untrimmed videos. The irrelevant Spatio-temporal features are detected in such conditions. To addressed this problem, Liu et al. [2] proposed a statistically prune Spatio-temporal features based methods to recognized and localization of action in untrimmed videos. They employed a PageRank method for informative static features mining. Finally, the heterogeneous features are fused and classified with AdaBoost classifier.

2.1.3.2 Local Features Approaches

After the recent development of visual recognition in static image numerous methods such as HoG, Bag of Features(BoFs) etc. are extended to video sequence analysis. In this context, Kläser et al. [62] proposed an action descriptor based on the histogram of oriented 3D Spatio-temporal features. They have extended and generalized HoG features based descriptors to 3D Spatio-temporal volumes for action recognition in videos. Furthermore, they extended the concept of integral images [63] to integral videos on gradient basis. For a given video sequence $v(x, y, t)$ the

integral video $v_{\partial x}$ along its partial derivatives in the x direction can be expressed as:

$$iv_{\partial x}(x, y, t) = \sum_{x' \leq x, y' \leq y, t' \leq t} v(x', y', t') \quad (2.3)$$

The integral video components $v_{\partial y}$, and $v_{\partial z}$, along y , and z are calculated Eq (1.3) accordingly.

A novel approach to automatically annotate the movie clips for training the action classifier, Laptev et al. [64] introduced the Histogram of Optical Flow (HOF) based spatiotemporal descriptor that is the extension of 2D Harris interest point detector to recognize the actions in realistic videos. Further, the bag of features based approach shown robustness to view-variations, illumination changes and clutter background conditions. Dalal et al. [65] developed a human pose descriptor using the Histogram of Oriented (HoG) to recognized the action in moving environmental conditions. This approach is fusing the gradient features with differential optical flow motion descriptor for the representation of human activities in realistic movie scenarios. The combined features of descriptor showed promising results in various challenging conditions. Kantorov and Laptev [66] proposed a reduced size video MPEG compression encoding technique for features extraction instead of optical flow that improved the speed of recognition around two times for representation of human actions. Further, the Fisher Vector encoding is used for action recognition and showed the efficient speed of operation with higher accuracy.

The BoWs feature based approaches [67] [68] are more dependent on segmentation of backgrounds, silhouettes extraction and optical flow estimation. It is found that such techniques are computationally complex and more prone to environmental disturbance. To overcome the existing problem in BoWs descriptors, Matikainen et al. [67] introduced ‘trajectons’ based on quantized trajectory tracked features. These ‘trajectons’ outperform the existing motion features descriptors for

recognition of human action in realistic scenarios. The spatiotemporal interest point extracted from cuboids are not so efficient as compared to trajectories of local features. Messing et al. [68] presented a generative mixture model based on the velocity history of tracked key points to recognize the human action in high-resolution video sequence. The proposed model extracted and tracked the features trajectories using Birchfield's implementation of KL tracker [69] that calculated the interest points where the eigenvalues are greater than a fixed threshold and tracked with consistency test by frame to frame translation. In [70] human actions and interaction recognized using trajectories features. However, these features are not so robust when camera motion and occlusions are present in videos, so sparse feature extraction is introduced for action detection [71]. The earlier sparse coding technique was useful for face recognition, image restoration, and subspace clustering.

In order to recognize the actions of different length and time scales in real-time a method based on the string kernel is proposed by Brun et al. [72]. They represented an action with the help of a string called 'aclet' and similarity between these 'aclets' is described based on Gaussian kernel.

2.1.3.3 Features Aggregation based Representation

In action recognition system, various action videos have not the same lengths. The various local Spatio-temporal features vector extracted are varying according to video lengths. In a machine learning algorithm such as SVM supports a fixed size vector length for a classification task. Therefore, to solve this problem of varying size features vector, we need a tool to aggregate sets of these local features descriptor into fixed-size vectors discriminative descriptor. In order to perform such task various dictionary learning [73] and Bag of Visual words(BoVs) [74] are introduced in the literature.

Zhu et al. [73] introduced a max-pooling sparse coding framework based on features extracted from dense Spatio-temporal features cuboids to recognize the human action in realistic unannotated videos. They adopted a HOG3D [62] descriptor to extract the dense Spatio-temporal features. Each extracted features are encoded sparsely with a pre-trained dictionary and max-pooling applied over the entire sparse code for each video. Guha and Ward [75] proposed a human action recognition sparse model based on the learned dictionary for video classification and behaviour analysis. The Spatio-temporal features descriptors are encoded and represented using sparse coding by over-complete learned dictionaries. However, this common dictionary approach is limited to when new action classes presented. Therefore, to alleviate this problem class-specific dictionaries are suggested to enhance accuracy.

Somasundaram et al. [76] proposed a dictionary learning approach based on global Spatio-temporal self-similarity score saliency to recognize the human action. Sadanand and Corso [77] introduced a high-level features representation model called 'action bank' for activity recognition in videos. These action banks were acting as a high-level dictionary and consisted of many low-level features descriptors obtained from semantic and viewpoint spaces. A relevant idea presented by Shao et al. [10] based on Laplacian of 3D Gaussian filter is used to represent the action space. Both of these methods based on pyramid structure to increase the robustness across the Spatio-temporal domains.

Gaidon et al. [78] presented an Actom Sequence Model (ASM) based on the temporal extension of the bag-of-features approach to recognize variable-length action videos. Actoms are formulated based on the sequence of atom units and the visual features are represented as a sequence of the histogram of Actoms. The initial success of Hidden Markov Models (HMMs) [79] in speech processing motivates the researcher for action recognition in videos because videos can be considered as a sequence of transition frames like state transition in HMMs. Hongeng and Nevatia

[80] proposed an approach using a modified semi HMMs model in conjunction with Bayesian networks for event detection in videos.

The inconsistency occurs due to first-order state transition improvised in the work of Tang et al. [81] using the max-margin variable interval HMM model. They utilized a conditional model in a trained max-margin structure that can discover the discriminative segments in the video simultaneously. Sun and Nevatia [82] addressed the problem of fixed size feature vector (e.g. BoW) by developing the concept of activity transition in video events. The Fisher Kernel methods are applied to facilitate the concept variables transitions over the interval can be encoded into a dense and fixed-size length feature vector.

Charletti et al. [83] proposed a hybrid feature extractors approach to recognize human activity using depth videos. The dimensionality of obtained feature vectors is reduced with the help of LDA and PCA techniques. Finally, the activities are classified using GMM classifier. However, they claimed superior accuracy but their model is sensitive to view variations and noises.

Ji et al. [27] introduced a soft-regression based transition maps approach for early detection of human activities using depth frames only. They divided human action into different patterns and evaluated the temporal coherence between action sequences. Inspired from the object recognition technique using hidden conditional random field (HCRF), Wang and Mori [84] presented a discriminative part-based approach called as max-margin hidden conditional random field (MMHCRF) using motion features for human action recognition in video sequences. Their model combined both large-scale global and local patch features to identify the different actions. Furthermore, a max-margin framework is used for learning the parameters of a hidden conditional random field model.

Liu et al. [85] proposed a hierarchical multi-task learning (HC-MTL) approach for joint human action grouping and recognition. They designed and jointly optimized the objective function into the group-wise least square loss regularized by

low rank and sparsely with respect to model parameters and grouping information. The non-convex optimization problem is solved by dividing the task into multi-task learning and task relatedness discovery. Xu et al. [86] proposed a two-layer hierarchical spatiotemporal model (HSTM) for recognition of complex human activities in videos. The HSTM model contains two hidden conditional random field layers in which the bottom layer used for spatial cues and top layer used for temporal information to characterize the video sequences.

Shan et al. [87] proposed a slicing representation based approach to recognize human action in videos. A minimum entropy method is used to select the optimal slicing angle for each video sequence, and then the slice sequences are converted into one-dimensional signals to represent the distribution of pixels along the time axis. Yu et al. [88] proposed a Gaze Encoding Attention Network (GEAN) based on the spatiotemporal sentence generation for video captioning in human activity dataset.

The et al. [89] presented a Spatio-temporal features approach based on Pachinko Allocation Model that describes the relation features for recognition of interactive activity. The intra and inter-person joint features of distance are calculated using the pose estimation outcome. The joint distance and angle features are represented by two codewords separately in the Poselet layer. The proposed model able to discriminate complex activities utilizing the correlation between generated body features and codebooks.

Unlike the Gait recognition in which human is identified only by walking activity, Yan [90] introduced discriminative sparse projections and ensemble learning-based approach for activity-based person recognition. The human body is projected into low dimensional subspace and collected into a number of clusters simultaneously. Yuan et al. [91] focused on action classification and annotations using temporal action localization in untrimmed videos. The Pyramid distribution fea-

ture (PSDF) model is proposed to extract motion cues in videos at multiple resolutions. Wang et al. [92] proposed a framework based on IDT features extraction and FV encoding for action spotting and localization in interaction videos.

Vishwakarma et al. [93] proposed a unified framework for based on the spatial distribution of edges gradient (SDEG) and R- transform to recognized human activity. Agahian et al. [94] proposed 3D skeleton joints based bag-of-poses a spatiotemporal model for action recognition. The K-means algorithm is used to train the pose descriptor and SVM to classify action pose.

Over the year, several handcrafted approaches introduced for action representation such as MEI, MHI, optical flow, MBH, HOG, HOF, Sparse representation, and dense trajectories. The most effective existing work for action recognized given by [8] [95] IDT with fisher vectors (FV). However, high computations complexity of IDT methods limits its implementation for real-time applications.

2.1.4 Still Image Based Action Recognition

We have reviewed previous works for human action recognition in still images and there are numerous spatial-temporal information-based techniques have been proposed for action recognition, but very few works are reported based on still images [96] [97] [98].

Thurau and Hlavac [99] introduced a human pose model feature descriptor for action recognition based histogram of the gradient (HOG) on a selected region of interest (ROI) and represented a feature vector using non-negative matrix factorisation. Raja et al. [100] proposed a subspaces graphical information approach using connecting images for human action recognition in still frames. Liu et al. [2] proposed an approach to recognized actions from the unconstrained realistic video. They utilized both motion and static feature and AdaBoost for final classification. In [101] they used non-negative matrix factorisation for high-level cues to recognise an action in still single images from a video frame of Weizmann

dataset, and Google downloaded images dataset. Zhang et al. [98] proposed a systematic approach to detect the shape of human interaction regions and a product quantisation approach was used for action labelling to obtain feature from the HOI parts. Zhao et al. [102] proposed the Riemannian projection model in which each video is considered as an image set and Grassmannian point is extracted for every six frames and projected through into a subspace using SVD.

2.1.5 Hand Crafted Feature Descriptors based on Skeleton Sequences

Seggesi et al. [103] proposed an automatic configurable trained feature extractor for the representation of skeleton pose from training samples data. However, their model showed promising results on static poses, but the response is found slow for real-time action recognition. To recognize the online action in a complex background using depth cameras, Ji et al. [26] proposed a hybrid approach by embedding skeleton coordinates into depth frames and extracts features using a spatiotemporal pyramid on a partitioned set of action sequences. However, these methods are showing satisfactory performance but less competent to tackle the environmental changes such as camera jitters, occlusion and illumination variations [65] [104] [57] [105].

Ghodsi et al. [106] proposed a spatio-temporal action template based on temporally averaging the action samples to recognize human activity using 3-D skeleton data. The actions classes are represented as multi-dimensional signals and able to deal with the variations in present in the activities such as speed. Yang et al. [107] introduced a latent max-margin multitasking learning approach for action recognition from skeleton data. Their skeletons model learned and correlate mid-level granularity of joints to represent action classes. Shabaninia et al. [108] proposed a weighted histogram 3D joints skeleton method to address the human activity recognition. Further, a weighted motion energy function is utilized to define the temporal variation and of actions.

The hybrid multimodal features approaches are introduced to compensate for the shortcomings of a single modality. A Gaussian descriptor is used to represent action and poses based on high order statistics of local features in two levels is proposed by Nguyen et al. [25]. They utilized K-means and sparse coding technique to compensate for the information loss generated by heterogeneous feature vectors obtained from two different input streams: depth and skeleton poses. Kong et al. [109] introduced a multi-modal fusion-based shared features learning approach to representing the local dynamics of each joint action class. The classification of action is performed using a max-margin framework. Raman and Maybank [110] presented a two-level hierarchical model based on Hidden Markov Model for activity recognition using depth sequences and skeleton coordinates. They provided a flexible representation to discriminates the action classes with similar activity labels.

Automatic Learned Feature-based Solutions

Deep learning approaches have been famous in industrial and commercial applications for a decade. Initially, these solutions are not performed well due to small datasets and hardware resources. With the advent of the latest computer technologies, it is possible to train these models over millions of videos samples. Presently, human action recognition with deep features extraction become trends in computer vision community due to the availability of larger videos datasets. Deep learning solutions let behind the handcrafted feature solutions as a comparison in robustness and classification accuracy.

The majority of deep learning architectures are based on CNN, and it only differs on how the input is applied. Few approaches using raw video frames as input while others using spatiotemporal features for training the network. The multitask deep learning and transfer learning approaches [32] [23] [111] were introduced, which combines the two datasets for action classification to availability medium datasets like UCF101, or HMDB51. The multiple task learning uses two-softmax

classification layer while in the transfer learning final fully connected layer is fine-tuned for a specific dataset.

In this sub-section, our objective is to discuss the deep learning architectures based approaches that have been introduced and used for learning action recognition, detection and localization in the videos in the past years. We have divided these deep models into four categories based on applied action classes as follows:

- Spatio Temporal Features based Architectures
- Multiple Stream Network-based Architectures
- Deep Generative Network-based Architectures
- Temporal Coherency network-based Architectures

In the following sub-sections, we have reviewed the above-mentioned architectures starting from earlier works to the state-of-the-art in details. Furthermore, we have discussed the merits and demerits of existing works.

2.1.6 Spatio-Temporal Features based Architectures

The first step towards the automatic features learning from raw RGB frames introduced by Ji et al. [5]. This approach is based on 3D ConvNets for human action recognition in videos. The 3D convolutional network extracts the spatial features along with the time axis; hence Spatio-temporal information is captured from the video for action representation. However, due to the rigid architecture of 3D CNN accepting the fixed input video frames (i.e. 7 frames), such model is not so efficient for action representation because different videos have a different time span.

To overcome the problem of a fixed number of input frames, Ng et al. [112] introduced ConvNets architectures based video level descriptor for recognition of action. The proposed architecture observed the effect of optical flow images with LSTM model and showed the robustness in terms of classification of action in

realistic videos. Karpathy et al. [32] introduced the concept of slow fusion in convolutional network layers to learn Spatio-temporal features. They proposed an architecture that processed the two streams, first is low resolution features context stream and second is high-level resolution fovea stream to reduce the time for training the CNN at promising speed. They showed that multi-resolution network improved the accuracy with the help of parameter sharing among layers.

Tran et al. [24] presented a 3D convolutional neural network (C3D) model based on the extraction of Spatio-temporal features for efficient video descriptors. They empirically showed that 3D ConvNet represented the temporal information in a better way as owing to 2D ConvNet features. Furthermore, the proposed C3D model is compact, simple and efficient for different video analysis. Donahue et al. [113] proposed a deep hybrid model that combines both CNN visual features and long range temporal recursion for human action recognition and video description. The Long Recurrent convolutional network (LRCN) extracts the visual features from varying length input videos using ConvNet and these extracted features are fed to long-short memory unit (LSTM) for extraction of sequential features. Both the CNN and LSTM architectures shared the parameters along time and resulting in a robust representation for action recognition.

Han et al. [114] proposed the dis-ordered multi-layer deep convolutional network, and they developed high-level features through transfer learning for action recognition in videos. Safaei and Foroosh [115] introduced a CNN model based on a prediction of the future motion of action in still images. They recognized shape and location feature in images with the help of the saliency map. Liu et al. [116] presented a view-invariant spatiotemporal deep network-based using skeleton joints for human activity recognition. Khaire et al. [28] proposed a three streams deep model to recognized human activity. They have constructed MHI from RGB, depth motion map sequences, and average skeleton images to trained the CNNs

model and extracted features are fused at decision level for final action classification.

It can be observed that shape and motion information are correlated in-depth sequences but are challenging to record both these simultaneously. Moreover, the 3D skeleton joint coordinates are not capable of discriminating some activities due to noise and occlusion errors such as self-occlusion with body parts etc. There are some action examples such as 'eating' and 'drinking' having the same motion pattern which cannot distinguish clearly using by 3D skeleton joints coordinates.

2.1.7 Multiple Stream Network-based Architectures

Simoyan and Zisserman [23] introduced two streams (spatial and temporal) network parallelly for action recognition. The spatial network receives raw input video frames while the temporal network accepts optical flow motion as input data. The spatial stream network fine-tuned to pre-trained network Imagenet while temporal stream trained using early fusion in the network on input optical flow fields. The temporal stream has multiple classification layers due to the limitation of a medium dataset, and each layer is trained separately, which leads to multiple task learning.

Feichtenhofer et al. [117] proposed a spatiotemporal multi-stream based ConvNet architecture for video action. They used multiplicative gating functions to utilise spatial and temporal information in the single forward pass. Baccouche et al. [4] introduced a hybrid model for action recognition that extracts spatial features using CNN and LSTM for motion cues.

Currently, a combination of both depth and local features are popular for activity recognition [8] [32] [95]. Feng et al. [49] proposed a geometrical relational features approach based on multilayer LSTM network for recognition of human activities using skeleton joints information.

Keçeli et al. [118] presented an approach to recognize the dyadic activity from depth sequence that is the combination of 3D and 2D CNN architectures. They extract the temporal features through 3D CNN trained 3D depth volume while 2D CNN is fine-tuned on weighted sum depth sequences. The obtained features are ranked using Reliff algorithm [118] and classify using an SVM classifier. Ijjina and Chalavadi [119] proposed a multimodal action recognition model based on feature extraction from RGB and depth videos using CNN architecture.

An ELM classifier is used to recognize the human activities from these fused features architecture. Jing et al. [120] presented a Spatio-temporal based hybrid neural network which characterizes human activity in RGB-D video based on a two-stream neural network. Further, they claimed to improvement in the classification accuracy by utilizing joint loss function to exploits the spatial and temporal features of videos.

Srihari et al. [121] introduced a four-stream CNN network consisted of two RGB-D video data and two temporal motion optical flow streams for recognition of human action. Elboushaki et al. [122] proposed a multi-dimensional CNN that learned high-level features for gesture recognition in RGB-D videos in conjunction with LSTM network. They investigated various fusion scheme at different layers of a deep network for the classification task.

2.1.8 Deep Generative Network-based Architectures

Deep learning generative model trains the temporal data in an unsupervised manner. It is a good option for video analysis because labelling the data is a difficult task and time-consuming. A deep generative model predicts future sequences accurately for motion dynamics.

Yan et al. [111] proposed a deep dynencoder to model the video dynamics that means study the characteristics between all pairs of adjacent frames in image sequences. A basic autoencoder and its variant are used to form the dynencoder followed by stacked strategy to deepen the architecture. The model is trained layer-wise pre-training and joint fine-tuning for understanding video dynamics.

Later on, to solve the constraint of smaller training samples, Srivastava et al. [123] introduced the multilayer LSTM network-based autoencoder model to predicted the future video sequences and reconstructed the input sequence. Goodfellow et al. [124] presented a discriminative adversarial network to overcome the difficulties present in the deep generative network. During training, this model learns with judgement whether the given input sequence is authenticated or not.

Mathieu et al. [125] proposed a multi-scale adversarial network for predicting the future sequence directly in pixel space. The proposed model addressed the problem sharpness in predicted future sequences by introducing an image gradients difference loss function to the realm the sharpness of the frames. They also discussed on merits and demerits of pooling technique in generative networks.

2.1.9 Temporal Coherency Network-based Architectures

It is observed that temporal annotation in each video sequence is a complex and time-consuming task. In the meantime, social sites such as YouTube produced numerous daily hour of untrimmed videos and annotated each video are impractical.

To solve the problem of annotation of untrimmed videos for action recognition and detection, Wang et al. [126] introduced an UntrimmedNet architecture for action recognition and detection in untrimmed video sequences. The proposed model implemented using end-to-end training and avoided the temporal annotations of action sequences. Cherian et al. [127] proposed a generalised low-rank pooling approach for action representation and video summarization that extracts

the features from intermediate CNN layers trained on subsequences. The conjugate gradient of the Grassmann manifold is for optimization of the proposed model.

Wang et al. [92] proposed view-invariant three-channel ConvNets architecture using weighted hierarchical depth maps for human action and interaction recognition. They projected spatiotemporal motion into the 2-D spatial structure. Mishra et al. [128] introduced the deep feature model based on temporal coherency to recognized action and posed. However, the temporal coherency based model is not a perfect choice for dynamic background videos. Lea et al. [129] introduced a time series model called temporal convolutional network (TCNs) that utilised hierarchical temporal convolution for recognition fine-grained human detection in videos. The proposed model decomposed into encoder and decoder temporal network used for pooling and up sampling respectively, for estimation the long-range temporal shapes efficiently.

Fernando et al. [130] presented a rank pooling function based method to extract video-wise temporal information for action recognition and video representation. For this objective, they assumed a video as a vector function that learned to order frames based appearance trajectories. The proposed approach is unsupervised based on temporal pooling that aggregates the information through the learning to rank procedure. Later on, Fernando and Gould [131] proposed end-to-end learning-based approach with backpropagation and temporal pooling mechanism for classification of video sequences. The proposed method coupled the CNN with temporal pooling layer that works on inner-optimization to encode the temporal semantics over long video clips into a fixed-length vector representation.

2.2 State-of-the-art Accuracy on RGB and RGB-D Datasets

Table 2.1 and 2.2 listed standard publically available RGB and RGB-D datasets. Furthermore, the highest recognition accuracies on these datasets obtained

through various state-of-the-art techniques and evaluation protocol are mentioned in tables. It is noted that the different feature descriptors are developed on these datasets in the past two decades, and no unique solution exists that can apply to both categories of RGB and RGB-D datasets. Earlier, action datasets recorded in a controlled environment with fewer numbers of samples and action classes. That is why they are more suitable for handcrafted features solutions such as in KTH, Weizmann, IXMAS, MSR Action 3D, CAD-60, Berkeley MHAD, and 50 salad datasets. Latest benchmarks recorded with multiple modalities in real scenarios such as UCF101, HMDB51, Hollywood, or YouTube 1M or 8M datasets.

Table 2.1: State-of-the-art Accuracy on RGB Dataset

Dataset	Classification Technique	Max Avg. Accuracy (%)	Evaluation Protocol	Year
Weizmann	EMRFs	100 [11]	LOOCV	2019
KTH	EMRFs	95.83 [11]	cross-validation	2019
IXMAS	HC-MTL+ L/S Reg	94.7 [85]	Cross-View	2017
CASIA Action	Hierarchical Spatio-Temporal model (HSTM)	95.24 [86]	-	2017
UIUC Sport	Adaptive Slicing feature, MFCC, SVM	98.9 [87]	LOSO	2015
Olympic Games	Motion Part Regularisation	92.3 [132]	LOOCV	2015
Hollywood	Joint max margin semantic features	48.58 [133]	-	2016
Hollywood2	(GEAN), RGP	92.40 [88]	-	2017
UT- Interaction	Hierarchical Spatio-Temporal Model	94.17 [86]	LOOCV	2017
UCF-YouTube	Bag of Expression(BoE)	96.68 [134]	-	2018
BEHAVE	(GIZ), (ARF + GCT + AF)	93.74 [135]	3-folds-cross-validation	2014

HMDB51	Spatiotemporal Multiplier Networks	72.20 [36]	-	2017
UCF50	ST-VLMPF(DF)	95.10 [136]	LOSO (Cross-View)	2017
BIT-Interaction	4-level , Pachinko Allocation Model	93 [89]	10-fold cross-validation	2016
MPII Cooking	GRP + IDT-FV	75.5 [92]	-	2015
UCF101	Spatiotemporal Multiplier Networks	94.9 [36]	-	2017
YouTube Sports 1M	HC-MTL+ L/S Reg	89.7 [85]	LOGO (Cross-View)	2017
ActivityNet	UntrimmedNet (hard)	91.3 [126]	-	2017
THU-MOS'15	Pyramid of Score Distribution Feature (PSDF)	40.9(0.1) [91]	-	2016
ChaLearn: Action/Interaction	Fisher vector + Idt features	53.85 [92]	LOOCV	2015
FCVID	Rdnn	76.0 [137]	-	2017
MOBISERV-AIIA	DSP+ OEML	68.0 [90]	-	2016
MERL Shopping	MSB-RNN	80.31 [138]	-	2016
YouTube 8M	NetVLAD + CG after pooling and MoE	83.0 [139]	-	2017
Okutama Action	SSD(RGB)	18.80 [140]	Cross validation	2017
20BN Something-Something (v1)	2D+3D-CNN (top-1)	63.80 [141]	-	2018
20BN Something Something(v2)	TPN(top-5)	91.28 [142]	-	2020

Table 2.2: State-of-the-art Accuracy on RGB-D(Depth) Dataset

Dataset	Classification Technique	Max Avg. Accuracy (%)	Evolution protocol	Year
i3DPost Multi-view	SDEG, + R transform	92.92 [93]	LOOCV	2016
MSR Action 3D	ConvNets	100	cross-subject	2015
RGB-D HuDaAct	BoW with χ^2 kernel SVM	82.9 [143]	cross subject validation	2014
CAD-60	HOG+BOW fusion	98.30 [144]	LOSO- CV	2020
50- Salads	ED-TCN	73.40 [129]	-	2017
Berkeley MHAD	Hierarchy of LDSs, HBRNN-L	100 [145]	-	2013
CAD-120	QQSTR with feature selection	95.2 [146]	4-fold cross validation	2015
Hollywood 3D	Bag of features (BoFs)	36.09 [147]	cross-validation	2014
MSR Action Pairs	Two stream coupled ConvNet	98.30 [148]	cross validation	2020
UWA3D Multi-view	MSO-SVM	91.79 [149]	cross-view	2015
Northwestern -UCLA	CNN+ Synthesized	92.3 [116]	cross-view	2017
LIRIS	Pose+ Appearance+ context	74 [150]	-	2014
IM-Daily Depth	ICP + KNN	81.17 [151]	cross- subject	2014
UTD-MHAD	Bag- of- Poses, ELM	95.0 [94]	LOSO- CV	2018
M ² I	FV/BoVs	92.33 [35]	cross -view	2017
SYSU- 3D HOI	SRNet	86.0 [152]	cross subject	2019
NTU RGB+D	SRNet	94.90 [152]	cross-view	2019

DAHLIA	Multi cam HOG	82.14 [153]	cross-subject	2017
PRAXIS Gesture	CNN + LSTM	84.74 [154]	-	2018
PKU-MMD	SRNet	97.10 [152]	cross-view	2019

2.3 Gaps Identified in the Present Study

Based on the literature survey, we have identified the following gaps:

- It is noted that global features based descriptors are too rigid to capture the possible variations (scale change, view variations, occlusion etc.) in the video frames. Furthermore, silhouettes feature extraction based solutions are not capable enough to find the fine details within the silhouettes.
- It can be visualized that abrupt scene change is a common phenomenon in videos. Some videos show little variation from frame to frame while other too fast. Therefore, it is not a good idea to extract key poses frames using normal distance function or some fixed threshold because of the risk of high information lost.
- Although the existing solutions are effective for action recognition from a similar viewpoint, the overall accuracy is degraded due to multi-view or view variations.
- It is observed that Convolutional Neural Networks(CNNs) based deep architectures shown the outstanding results, for understanding the image content and video-based analysis. From a computational viewpoint, these models need additional time and larger dataset samples to train the model effectively and optimize the millions of parameters for video representations.
- It is observed that optical flow-based solutions are less effective in challenging conditions such as view-variations and occlusions present in videos.
- It is observed that the traditional feedforward network and convolutional neural network are inefficient to deal with sequential data such as video

analysis. The RNN deep network has capability to deal with sequential data. However, RNNs are inefficient for longer duration sequential input because of the vanishing gradient in the backpropagation process.

2.4 0. Research Objectives

A robust human action recognition system must be generic, compact, efficient, and straightforward. To overcome the limitations in at least one of aspects such as similarity of actions, cluttered background, viewpoints variations, illuminations variations, and occlusions, the main objectives of this thesis work are as follows:

- To review state-of-the-art handcrafted and deep learning approaches, standard single and multimodal human activity datasets, existing solutions and their limitations.
- To develop an effective approach for representation of various human activities.
- To develop an algorithm which can handle issues of low illumination, noise, cluttered background and environmental conditions.
- To develop an algorithm which can enhance the video quality so that representation is more efficient than available resources.
- To develop a robust algorithm to classify the represented activity.
- To validate the developed algorithm and experiment can be conducted on standard datasets.
- To know the effectiveness of the novel algorithm a comparative study and implementation is to be conducted.
- To develop useful feature representation and classification using a deep learning based model simultaneously.

Chapter 3

Human Activity Recognition Using Hand-crafted Features

This chapter presents a novel hand-crafted features descriptor for human action recognition in video sequences. This method addressed the major challenges such as abrupt scene change phenomena, clutter background and viewpoints variations by presented a novel visual cognizance based multi-resolution descriptor for action recognition using key pose frames. The proposed framework is constructed by computation of textural and spatial cues at multi-resolution in still images obtained from videos sequences, which is known as Extended Multi-Resolution Features (EMRFs) model. The effectiveness of the proposed approach is explained and validated through experiments on standard datasets and state-of-the-art comparison of obtained results.

3.1 Introduction

The objectives of this chapter are to develop a novel handcrafted feature-based descriptor for human action recognition in video sequences. Recently, action recognition in still frames become an imperative choice for a researcher in computer vision. The still image-based recognition tries to find out a person's behaviour or action using only a single image. Although, most of the present literature regarding behaviour analysis uses videos analysis in which temporal and spatial cues are used. It can be considered more challenging to recognised action in still image than video analysis because it does not involve the temporal variations, illumination variation and alignment of the images.

3.2 A Visual Cognizance Based Multi-Resolution Descriptor for Human Action Recognition using Key Pose

In this framework, a robust architecture is constructed by computation of textural and spatial cues at multi-resolution in still images obtained from videos sequences. A fuzzy inference model is used to select the single key pose image from action video sequences using maximum histogram distance between stacks of frames. To represent these key pose images, the textural traits at various orientations and scales are extracted using Gabor wavelet while shape traits are computed through a multilevel approach called Spatial Edge Distribution of Gradients (SEDGs).

Hence, a hybrid model of action descriptor is developed using shape and textural evidence, which is known as Extended Multi-Resolution Features (EMRFs) model. The action classification is carried through two most famous and efficient distinctive classifiers known as SVM and k-NN and compared individual recognition accuracies. The developed model showed the outstanding results on standard human action datasets with SVM classifier. The performance of the proposed framework shows supremacy as compared with earlier state-of-the-art approaches. The underlying architecture of the proposed work is depicted in Fig. 3.1.

3.2.1 Selection of Single Key Pose using Fuzzy Logic

We extract the stacks of key poses frames from input action video sequences using histogram distance between adjacent frames. Further, we select the single key still image from these stacks of frames based on Fuzzy logic inference model. The details descriptions of each block are explained in the following sub-sections.

3.2.1.1 Key Poses extraction from Action Video Sequence

Human activity can be effectively recognized with the help of key poses extracted from the video sequence and provides an explicit representation of human body

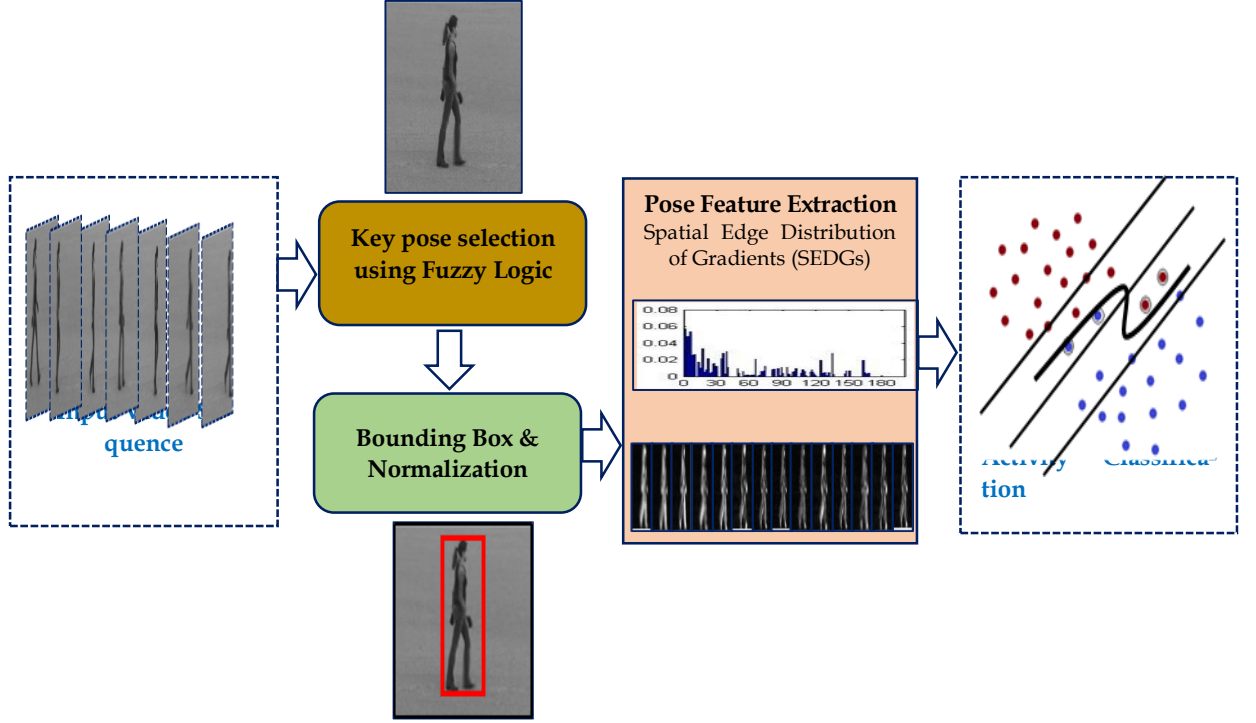


Figure 3.1: Shown Visual Cognizance Based Multi-Resolution Descriptor for Human Action Recognition using Key Pose

posture. These key poses are chosen as a single prime keyframe by using histogram distance which shows a spatial representation of 2D posture of human body motion. The key pose frames are selected as stacks of ten frames at a regular interval from the input video sequences. To make these stacks of frames device invariant, they are transformed into a CIE Lab colour space [155]. The histogram distances are calculated for three different parameters, Luminance(L), and hue angle axis(a) at 0° and hue angle(b) at 90° of CIE Lab color space. The histogram distance (\mathcal{D}) between the frames can be calculated and given by using Eq. 3.1.

$$\mathcal{D} = \left\| \sum_{i=1}^M \sum_{j=1}^N S_{ij}^t - S_{ij}^{t+1} \right\| \quad (3.1)$$

where, S is the stacks of frames of size $M \times N$ and t represents the number of frames. The histogram distance is useful for measure instance changes in the frames of input video sequences. The computed distances for adjacent frames are utilized for the selection of key poses in the following sub-section.

3.2.1.2 Fuzzy Inference model

It can be visualized that abrupt scene change is a common phenomenon in videos. Some videos sequence shows little variation from frame to frame while other too fast. It is not a good idea to extract key poses frames using normal distance function or some fixed threshold because of the risk of high information lost. Therefore, a fuzzy logic rule-based model is utilized to extract key pose images from video sequences. Further, the histogram distances are calculated for adjacent frames for selecting optimised keyframes.

The fuzzy logic models proposed by L. Zadeh [156] is based on the degree of membership. This theory is the extension of Crips Boolean logic which takes a hard decision whether a particular class belongs to a group or not such that:

$$\mu_A(x) = \begin{cases} 1 & \text{if and only if } x \in A \\ 0 & \text{if and only if } x \notin A \end{cases} \quad (3.2)$$

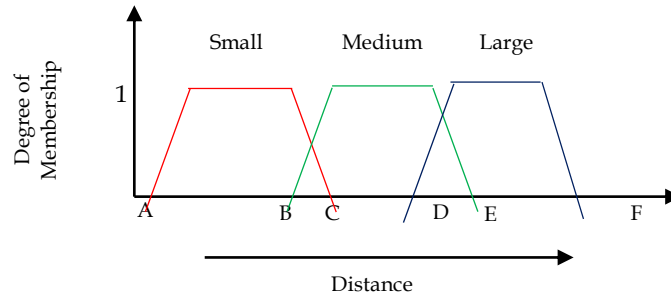


Figure 3.2: Fuzzy trapezoidal membership function for selecting the key poses frames

On the other hand, Fuzzy logic is simple and deals with rule-based IF X AND Y THEN Z approach to assign the degree of membership to a particular class rather than to model a system mathematically. Fuzzy logic assigns a flexible membership between 0 and 1 to a variable class for a particular group. Fig.3.2 depicts the Fuzzy trapezoidal membership function which is used for selecting the keyframes from video sequences.

In this approach, the keyframes are selected using the fuzzy model, as shown in Fig. 3.2. These key pose frames are internally compared and ranked according to histogram distance. It is found that the higher value of histogram distance frames shows maximum variations as compared to other frames. If the extracted key frames are denoted as $f_1 f_2 f_3 f_4 \dots f_n$, then the histogram distance for adjacent frames is calculated as:

$$key\ pose\ frame = \arg \max(D_i), \quad i \in n \quad (3.3)$$

The highest distance keyframes that is having higher pixel difference as compared with other and selected as a single key still frame as illustrated in Fig. 3.3. The proposed algorithm for single key pose extraction is listed in Algorithm 1.

Algorithm 1: Fuzzy Inference Model for Selection of Single Key Pose

Step 1: The image frames are selected at the interval of 10 frames from the input video sequence.

Step 2: Selected frames are converted into new colour space 'CIELab'.

Step 3: Histogram distances are computed for ' L '(**lightness**), ' b '(**hue angle axis at 90°**) and ' a '(**hue angle axis at 0°**), parameters using Eq.(3.1)

Step 4: The means u_b are computed for all adjacent frames differences as:

$$u_b = \left\lfloor \frac{Count\ L + Count\ b + Count\ a}{3} \right\rfloor$$

Step 5: Find the values of endpoints components of the membership function shown as in Fig.3.2 $A = (u_b - u_b * 0.4)$, $B = (u_b - u_b * 0.3)$, $C = (u_b - u_b * 0.2)$, $D = (u_b + u_b * 0.4)$, $E = (u_b + u_b * 0.5)$, $F = (u_b + u_b * 0.8)$.

Step 6: Create the trapezoidal fuzzy membership function for computed mean u_b as depicted in Fig. 3.2, where linguistic parameters are defined as: small, medium, and large.

- Rule 1: IF the distance between a segment frame and its neighboring segment frame is “medium” THEN it is a key frame.
- Rule 2: IF the distance between a segment frame and its neighbouring segment frame is “large” THEN it is a key frame.
- Rule 3: IF the distance between a segment frame and its neighboring segment frame is “small” THEN it is NOT a key frame.

Step 7: Set the fuzzy rules based on neighbouring small or large distances frames to extract key pose frames.

Step 8: Compared to the selected frames with internal frames and ranked according to a histogram distances difference between them is a single still key pose and denoted as $f_s(x, y)$ of size $M \times N$ as depicted in Fig. 3.3.

3.3 Extended Multi-Resolution Features (EMRFs)

We can observe that human actions are visualized and defined by the movement of different body postures. The movement of these body postures reflects the visual cues in the video scene and help to distinguished the actions. The proposed model is based on human visual perceptions that extract the shape and orientations features are from still pose images to recognize the action class. The shape features are using Spatial Edge Distribution of Gradients (SEDGs) and orientation features at different scales are represented by Gabor wavelet transform for action classification. The detail explanation of each features representation is discussed in the following subsections:

3.3.1 SEDGs Feature Map

The body posture performed by human contains information about body motion. The 2-D representation of these images gives spatial information about human motion. Such spatial distribution of postures provides the attitude of action

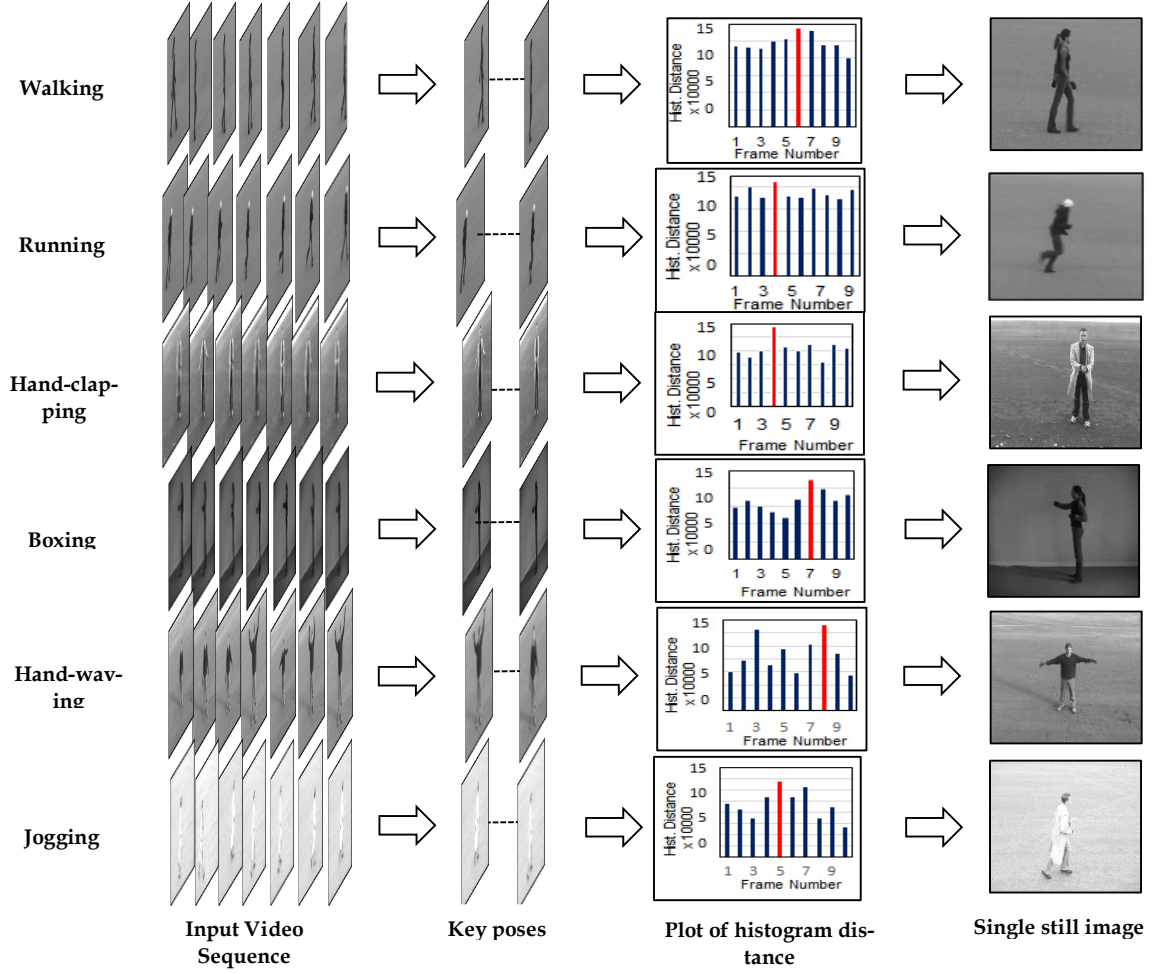


Figure 3.3: Illustration of workflow for selecting a single image from input video of KTH dataset

behaviour of persons. The still image approaches are less complex and time efficient. Edges detection are the most important task for efficient shape feature extraction from an image. In our approach, we have efficiently utilized the Canny edge detector [157] to obtain the edges and a threshold mechanism based on pixel variation are employed to remove unnecessary edges present in an image. To extract shape feature, the region of interest (ROI) is chosen which further divide into sub-regions at different sub-levels. The proposed algorithm of SEDGs is shown in Algorithm 2.

Algorithm 2: SEDGs Feature Map

Step1: Select 't' set of frames from input video sequences $F(\mathbf{x}, \mathbf{y}, \mathbf{t})$.

Step 2: Select a single key pose frame from as described in section 3.2.

Step 3: Choose ROI and normalized it to the fixed spatial dimension 50×50 , and denoted as:

$$\mathcal{R}(\mathbf{x}, \mathbf{y}, \phi), \text{ for all } 0 \leq \mathbf{x}, \mathbf{y} \leq 50.$$

Step 4: Apply the canny edge detector to detect edges of selected ROI and given as:

$$\epsilon(\mathbf{x}, \mathbf{y}, \phi) = \text{canny}(\mathcal{R}(\mathbf{x}, \mathbf{y}, \phi))$$

Step 5: Find the spatial edge distribution vector at any point (\mathbf{x}, \mathbf{y}) of the entire image $\epsilon(\mathbf{x}, \mathbf{y}, \phi)$ different sub levels as:

- i. At level-0, the Magnitude $M(\mathbf{x}, \mathbf{y}) = \sqrt{\mathfrak{N}_x(\mathbf{x}, \mathbf{y})^2 + \mathfrak{N}_y(\mathbf{x}, \mathbf{y})^2}$ and Orientation $\phi(\mathbf{x}, \mathbf{y}) = \arctan \frac{\mathfrak{N}_x(\mathbf{x}, \mathbf{y})}{\mathfrak{N}_y(\mathbf{x}, \mathbf{y})}$. Where $\mathfrak{N}_x(\mathbf{x}, \mathbf{y})$, $\mathfrak{N}_y(\mathbf{x}, \mathbf{y})$ are \mathbf{x} and \mathbf{y} direction gradients of image respectively. Each sub region is quantized into the 8 orientation bins evenly distributed between 0° to 360° . The resulted feature vector for selected ROI is of 8×1 dimension.
- ii. At level-1, the total region of an image $\epsilon(\mathbf{x}, \mathbf{y}, \phi)$ is sub-divided into 4 sub-image regions, and represented as: $\epsilon(\mathbf{x}, \mathbf{y}, \phi) = \{\mathcal{S}_1(\mathbf{x}, \mathbf{y}, \phi), \mathcal{S}_2(\mathbf{x}, \mathbf{y}, \phi), \mathcal{S}_3(\mathbf{x}, \mathbf{y}, \phi), \mathcal{S}_4(\mathbf{x}, \mathbf{y}, \phi)\}$. The feature vector of dimension $8 \times [1 + 4]$ is formed using (step 5-i).
- iii. At level-2, Each of these sub-blocks (Step 5-ii) are further divided into four sub-blocks. A feature vector is of dimension $8 \times [1 + 4 + 16]$ is obtained from 16 sub-block as in (step 5-i).

Step 6: A final feature vector based on spatial edge distribution is formed and summing all vectors of sub-levels and represented as: $\mathcal{F}_\epsilon = 8 \times [1] + 8 \times [1 + 4] + 8 \times [1 + 4 + 16] = 216$. The results of the final feature vector are depicted in Fig. 3.4.

Fig. 3.4 and 3.5 depict the results obtained at a different level using SEDGs feature extractor. It can be observed from Fig. 3.5 that the extracted shape features are

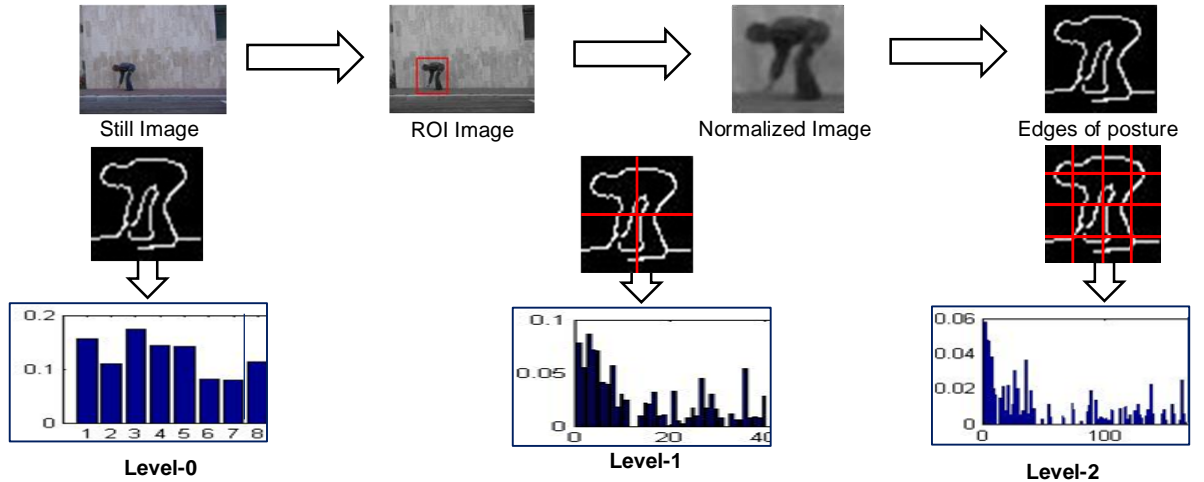


Figure 3.4: Simulation Results of Proposed Algorithm 2

discriminative due to the variation of the histogram for different activities. Therefore, these features are robust to represents human activities. It is a fast and straightforward approach to calculate the features based on spatial shape information.

3.3.2 Orientation Feature Map

The orientation information of the action pose is extracted by Gabor filter, which is one of the widely used techniques for orientation and texture in the image. Arivazhagan et al. [158] introduced an invariant rotation approach for texture classification based on wavelet and defined as per Eq. (3.4).

$$\psi(x, y) = (2\pi\sigma_x\sigma_y)^{-1} \exp\left(-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)\right) \exp(2\pi j\omega x) \quad (3.4)$$

where σ, ω are scale and modulation frequency respectively. For a given still image $f_s(x, y)$ of size $M \times N$, the Gabor wavelet transform (GWT) at scale p and orientation q is obtained by convolving $f_s(x, y)$ with $\psi_{pq}^*(x, y)$ as per Eq. (3.5).

$$\mathcal{G}_{pq} = \sum \sum f_s(x, y) * \psi_{pq}^*(x, y) \quad (3.5)$$

where ψ_{pq}^* is complex conjugate of mother wavelet as given in Eq.(4). The Gabor wavelets are formed by exciting function as given in Eq. (3.6).

$$\psi(x, y) = a^{-p} \psi(X, Y) \quad (3.6)$$

where $X = a^{-p}(x\cos\theta + y\sin\theta)$, $Y = a^{-p}(x\cos\theta - y\sin\theta)$, θ is orientation parameter, a is the scaling parameter, for $a > 1$, $\theta = \frac{q\pi}{Q}$. $p = 0, 1 \dots \mathcal{P} - 1$, $q = 0, 1 \dots \mathcal{Q} - 1$. \mathcal{P} and \mathcal{Q} are the total number of scale and orientations respectively.

The still image at various scales and orientations can be represented through the convolution of the image with Gabor wavelet transformed as shown in Fig.3.6, which have three scale and eight orientations. The obtained images at different orientations and scales are arranged according to the energy content. The orientation with the highest energy image is called the dominant orientations.

Hence, the feature extracted from these images has to place first in the feature vector. The energy is computed using Eq. 3.7.

$$u(p, q) = \sum_x \sum_y \|\mathcal{G}(x, y)\| \quad (3.7)$$

The mean and standard deviation of all the transformed coefficient is computed using Eq. 3.8. These values represent the region of homogenous texture in the image.

$$\mu_{pq} = \frac{u(p, q)}{MN} \quad \text{and} \quad \sigma_{pq} = \sqrt{\frac{\sum_x \sum_y \|\mathcal{G}(x, y)\| - \mu_{pq}}{MN}} \quad (3.8)$$

A Gabor feature map \mathcal{F}_G are formed for \mathcal{P} : Scales and \mathcal{Q} : Oreinations as in Eq.3.9.

$$\mathcal{F}_G = [\mu_{00}, \sigma_{00}, \mu_{01}, \sigma_{01}, \mu_{02}, \sigma_{02}, \dots \dots \dots, \mu_{p-1q-1}, \sigma_{p-1q-1}] \quad (3.9)$$

3.4 EMRFs Representation

The EMRFs model is inspired by the human visual cognizance [159] as shown in Fig. 3.7. The EMRFs representation is achieved by concatenated the features obtained from SEDGs and GWTs vectors. The proposed model provides an informative representation of the human activities by capturing shape and

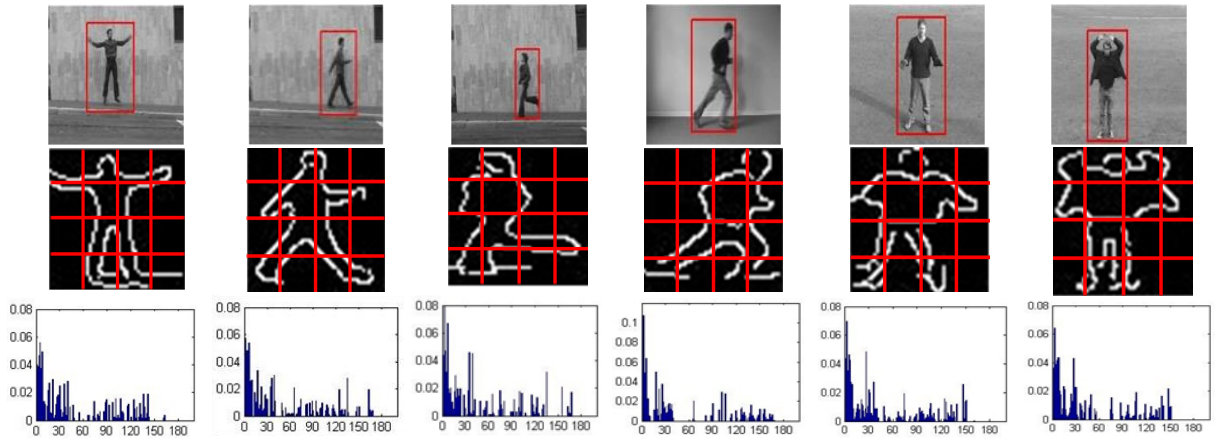


Figure 3.5: First row shows Region of Interest (ROI) on various activities images, Second row shows edges computed on different postures, and third row represents histogram of SEDGs at level '2'.

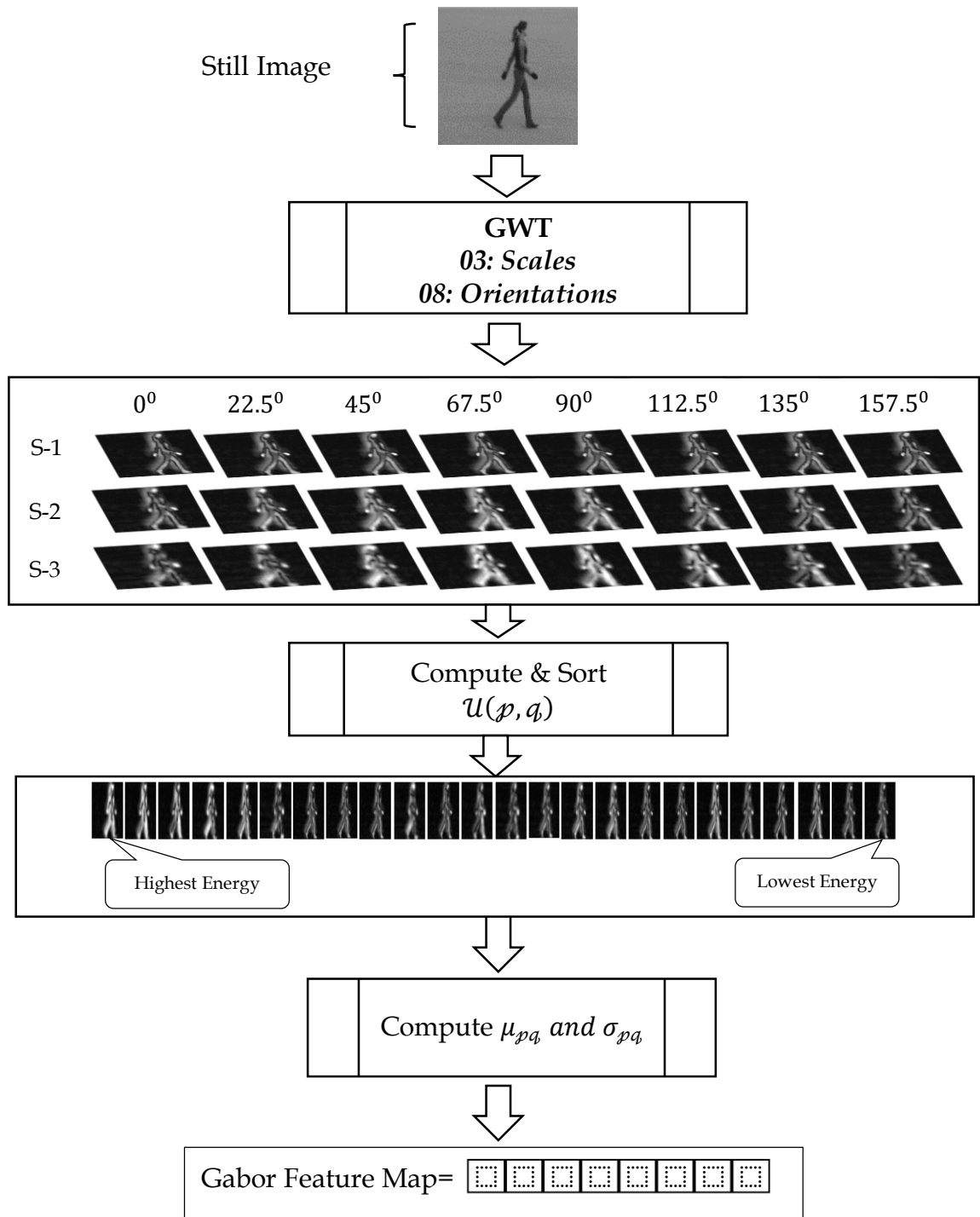


Figure 3.6: Shown the orientation features extraction vector map using Gabor Wavelet Transform

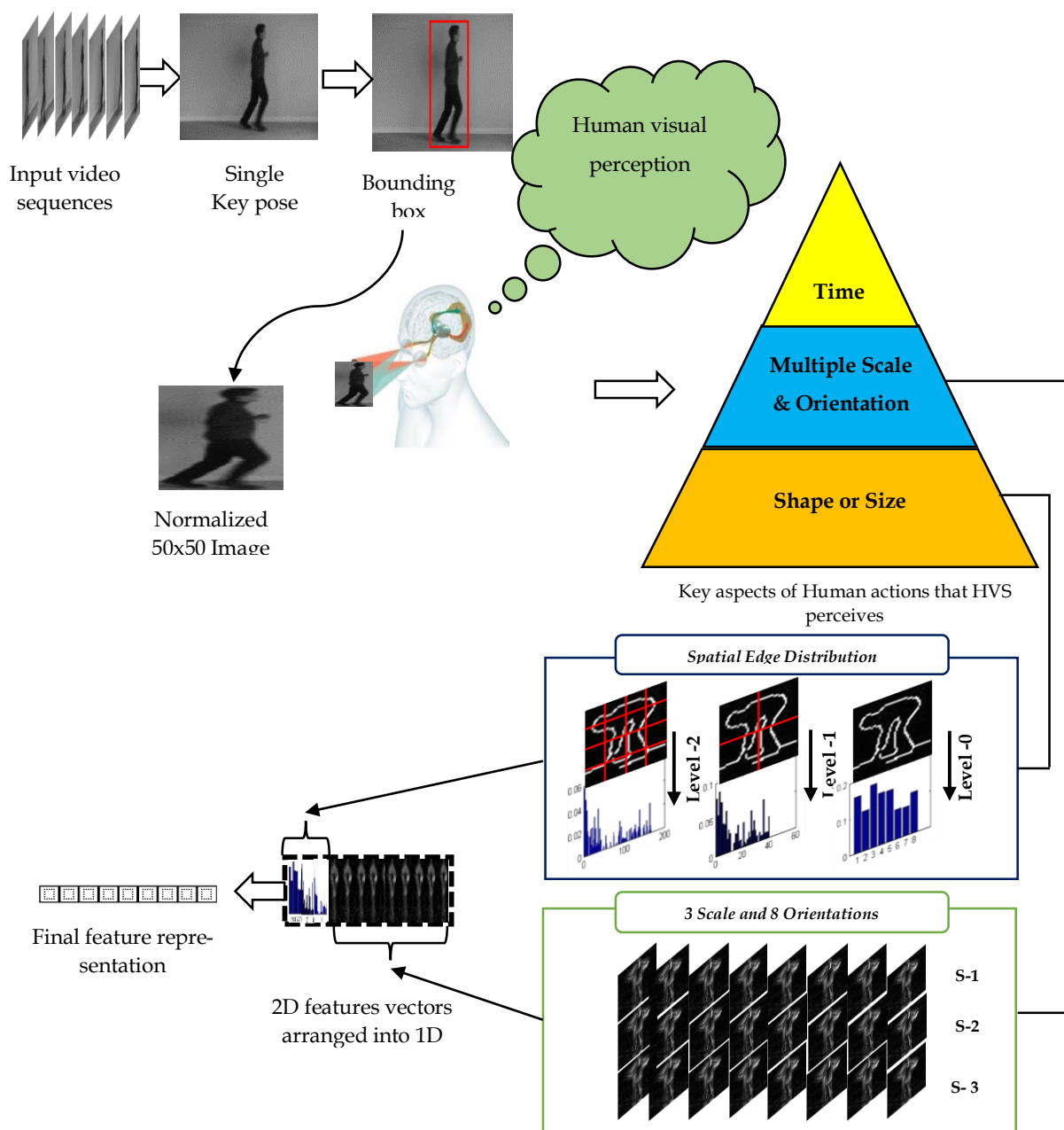


Figure 3.7: Procedure of proposed EMRFs framework for HAR

orientations features at multiple resolutions, scales and bins. Hence, the EMRF is robust enough to deal with challenging situations such as translation, rotations, and scaling, rotation. The effectiveness of EMRFs representation is measured by experimenting on human actions datasets.

3.4.1 Performance Evaluation on Human Action Datasets

In this section, the performance of our proposed approach is evaluated on four publically available human action datasets such as Weizmann Action [54], KTH [29], Ballet Movement [160] and UCF YouTube [2]. In the following sub-sections, a brief description of all four datasets and different experimental conditions are presented.

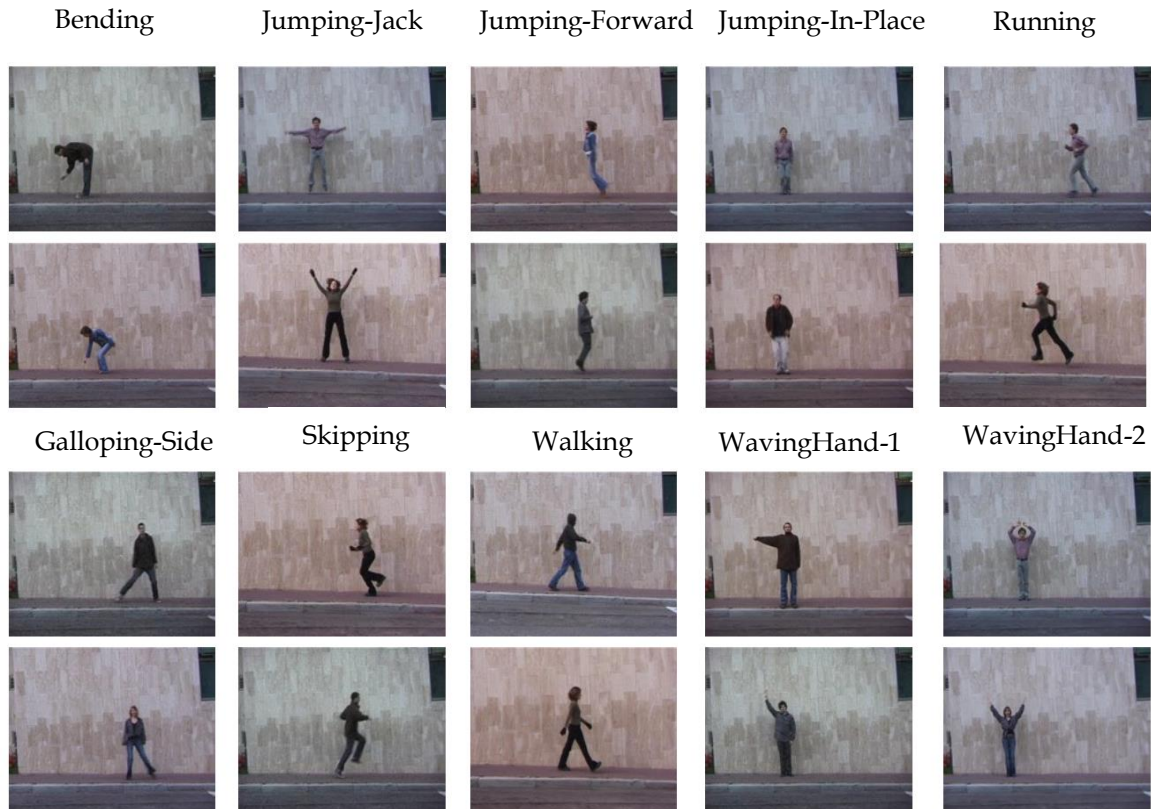


Figure 3.8: Example frames from Weizmann Action Dataset

3.4.2 Weizmann Action Dataset

Blank et al. [54] introduced this dataset for simple atomic action recognition in videos. It is recorded with normal RGB cameras in outdoor environmental conditions. It consisted of 10 action classes performed by 9 people such as: Run, Side, Skips, Jump, jumping -jack, Bend, Jack, Walk, Wave-1, and Wave-2. There are a total of 90 videos clips recorded at 15 fps with 144×180-pixel spatial resolution. The examples frames from Weizmann action dataset are shown in the Fig.3.8.

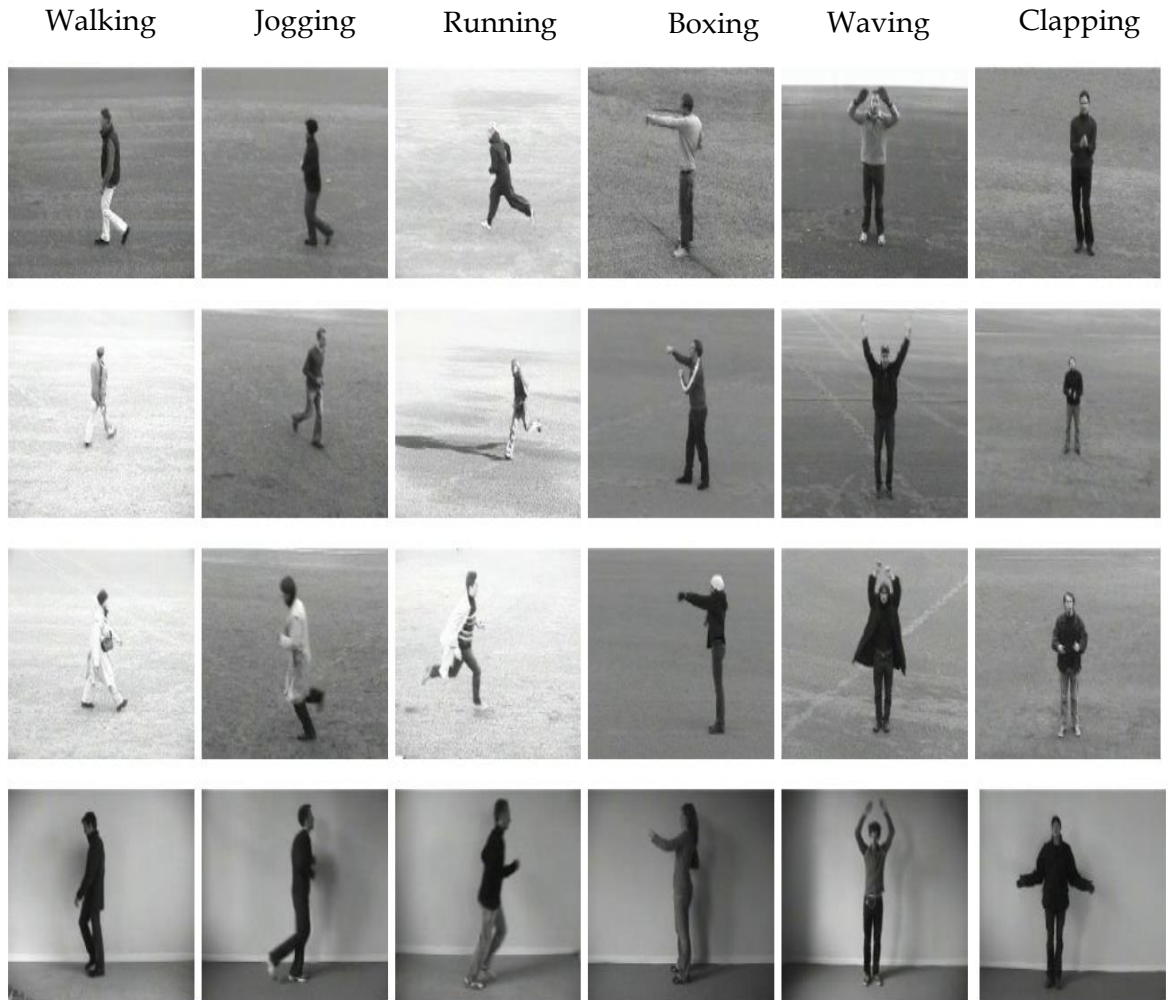


Figure 3.9: Example frames of KTH Action Dataset

3.4.3 KTH Action Dataset

KTH action dataset [29] is the most widely used dataset for human action analysis.

It is more challenging action dataset as compared with Weizmann because changing illumination conditions and outdoor environments. It consists of 25 subjects performing six activities such as jumping, hand-clapping, jogging, walking, hand-waving, and running. There is a total of 600 videos recorded with 25 fps with spatial resolution 160×120 pixels, in 4 different scenarios. The example frames from KTH action dataset is depicted in Figure 3.9.

3.4.4 The Ballet Dataset

Fathi and Mori [160] collected this dataset from ballet DVD videos. It consists of 8 movement activities performed by 1 woman and 2 men actors as: turning(TR), jumping(JP), left to the right-hand opening(LRHO), standing still(SS), right to the left-hand opening(RLHO), leg swinging(LS), hopping(HP), and stand with hand opening(SWHO). There are total of four annotated video sequences in each video. This dataset is challenging regarding large intra-class dissimilarity and inter-class similarity, clothing variations, the speed of activity, and spatiotemporal variations. The examples frames from ballet movement dataset are shown in Fig. 3.10.

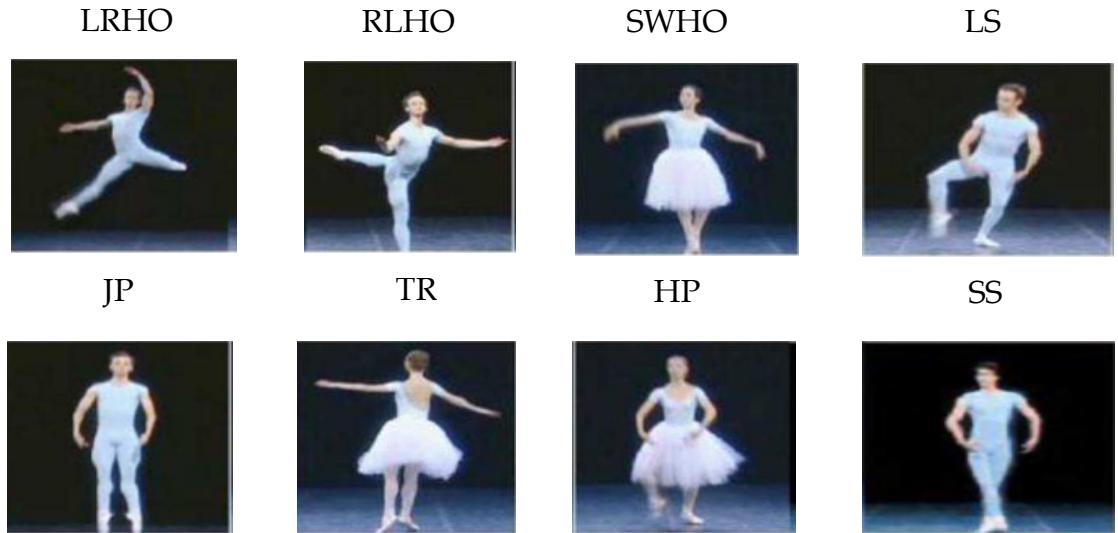


Figure 3.10: Example frames from Ballet Movement Dataset



Figure 3.11: Example frames from UCF YouTube Action Dataset

3.4.5 UCF YouTube Action

Liu et al. [2] introduced this dataset to recognised complex action in a realistic environment. It consists of 11 action classes as tennis swinging, volleyball spiking, soccer juggling, walking with a dog, basketball shooting, cycling, horseback riding, swinging, golf swinging, trampoline jumping, and diving. There are twenty-five video groups each having four video clips sharing common features. These dataset videos are challenging due to view-points, occlusions, varying illumination conditions, and clutter backgrounds. The samples image frames from UCF YouTube dataset is shown in Fig. 3.11.

3.5 Experimental details and Results Analysis

In this section, we discussed the experimental setup evaluations protocols applied for different human action datasets. Since the developed approach is computed on

still key pose image, hence, where the action is captured in the video, first the key pose are images are extracted from the video as explained in sections 3.2 and 3.3 respectively. The spatial distribution features map is divided into eight orientations bins in the range of $[0, 360^0]$. On the other hand, Gabor features map is calculated with three scales and eight orientation angles. Finally, action recognition is done by combining the extracting the spatial and textural features vectors maps. The action classification is carried through a linear Support Vector Machine (SVM) [161] and a non-parametric technique K-nearest neighbour (k-NN) [162] classifiers. The accuracy of our EMRFs is measured as average recognition accuracy (ARA) in leave-one-out-cross-validation (LOOCV) evaluation protocol. The ARA is calculated using Eq. 3.10.

$$\text{Accuracy} = \frac{\text{True Postive} + \text{True Negative}}{\text{True Postive} + \text{TrueNegative} + \text{False Positive} + \text{FalsePositive}} \times 100\% \quad (3.10)$$

The classification accuracy is computed by dividing the correct number of predicted images by the number of tested images. The classification accuracies in the form of confusion matrices using k-NN and SVM classifier on four datasets are shown in Fig. 3.12 and 3.13, respectively.

The confusion matrix obtained from k-NN classifier for all dataset is shown in Fig. 3.12. It can be observed from the confusion matrix of Weizmann dataset depicted in Fig.3.12(a) that the accuracy achieved on this dataset is very high. Similarly, the confusion matrixes of KTH, Ballet, and UCF YouTube datasets are shown in Fig.3.12(b), (c) and (d) respectively. The activities in other video datasets are classified with fewer ambiguities. The accuracy is achieved on Ballet dataset is less as compared with other datasets because of existing challenges such as self-occlusion and high intra-class similarity of actions. It can be inferred from the dataset too.

The classification accuracies in the form of a confusion matrix with SVM classifier for all datasets is shown in Fig. 3.13. In the presence of various challenges

in the videos such as high interclass similarity in still key poses of running, walking, and jumping. Besides this our model performed well and most of the activities are discriminated with the highest recognition accuracy. The confusion matrix of the Weizmann dataset is shown in Fig. 3.13(a). There is slightly less confusion in the case of 'jack hand' and 'wave-2' due to similar key poses. In Fig. 3.13(b) the confusion matrix for KTH dataset shows much satisfactorily results with average recognition accuracy 95.85 % and very less misclassification is found only three classes: 'running', 'walking', and 'jogging' due to similar key still key poses. Further, our model efficiently classified with maximum accuracy on the rest three action classes. The ARA of 92.75 % is obtained on Ballet dataset shown in the confusion matrix in Fig. 3.13(c). Although action recognition in this dataset is complex due to clothing, gender, and size variations. But the proposed EMRFs model is insensitive to these variations and complexity of actions. It can be observed from the confusion matrix Fig.3.13(c) that there are little bit confused about still key pose of action pair such as 'hopping' and 'jumping', 'leg swing' and 'Right to left-hand opening', 'turning right' and 'stand hand opening', and 'jumping and standing still' beside this our model gives comparable accuracy with the existing state-of-the-art methods [114], [163]. The confusion matrix of UCF YouTube dataset is shown in Fig. 3.13(d). Some similar action is creating the confusion in motion feature such as 'cycling' and 'horseback riding' and 'jogging' and 'running'. Our method gives the highest accuracy on this challenging realistic video dataset.

The accuracy achieved through k-NN and SVM has been compared as shown in Fig. 3.14. It can be seen from the figure that the performance of SVM is better than k-NN for all dataset because the k-NN classifier shows best results with higher dimensional training data and a larger value of k, but the large values of k lead to high computation. However, the ARA is varying from one data set to another because of the recording conditions and environment setting of the dataset.

The highest accuracy achieved for all datasets through SVM has been compared with the similar state-of-the-art.

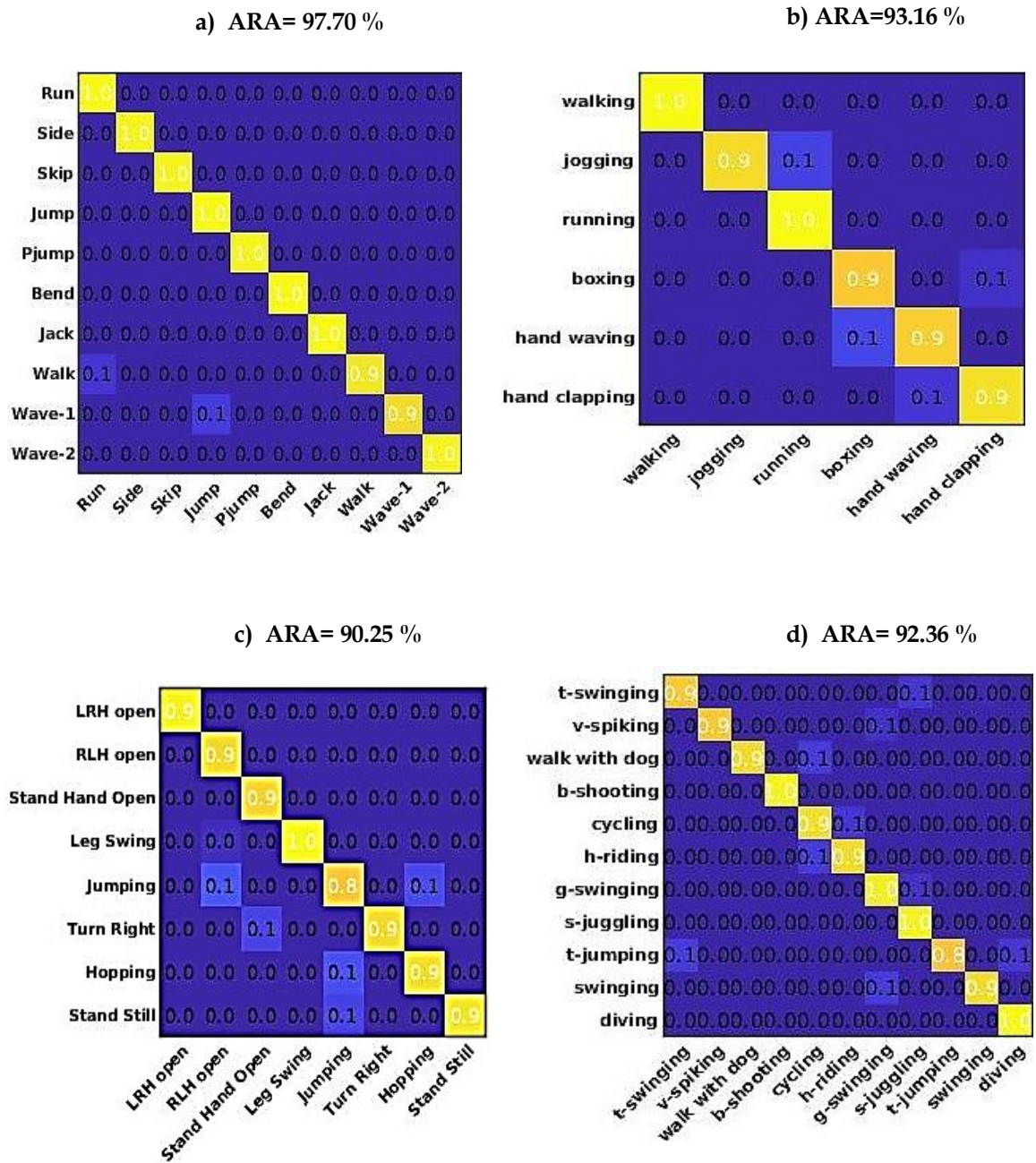


Figure 3.12: Classification result of k-NN classifier on (a) Weizmann Action (b) KTH (c) Ballet Movement (d) UCF YouTube action datasets.

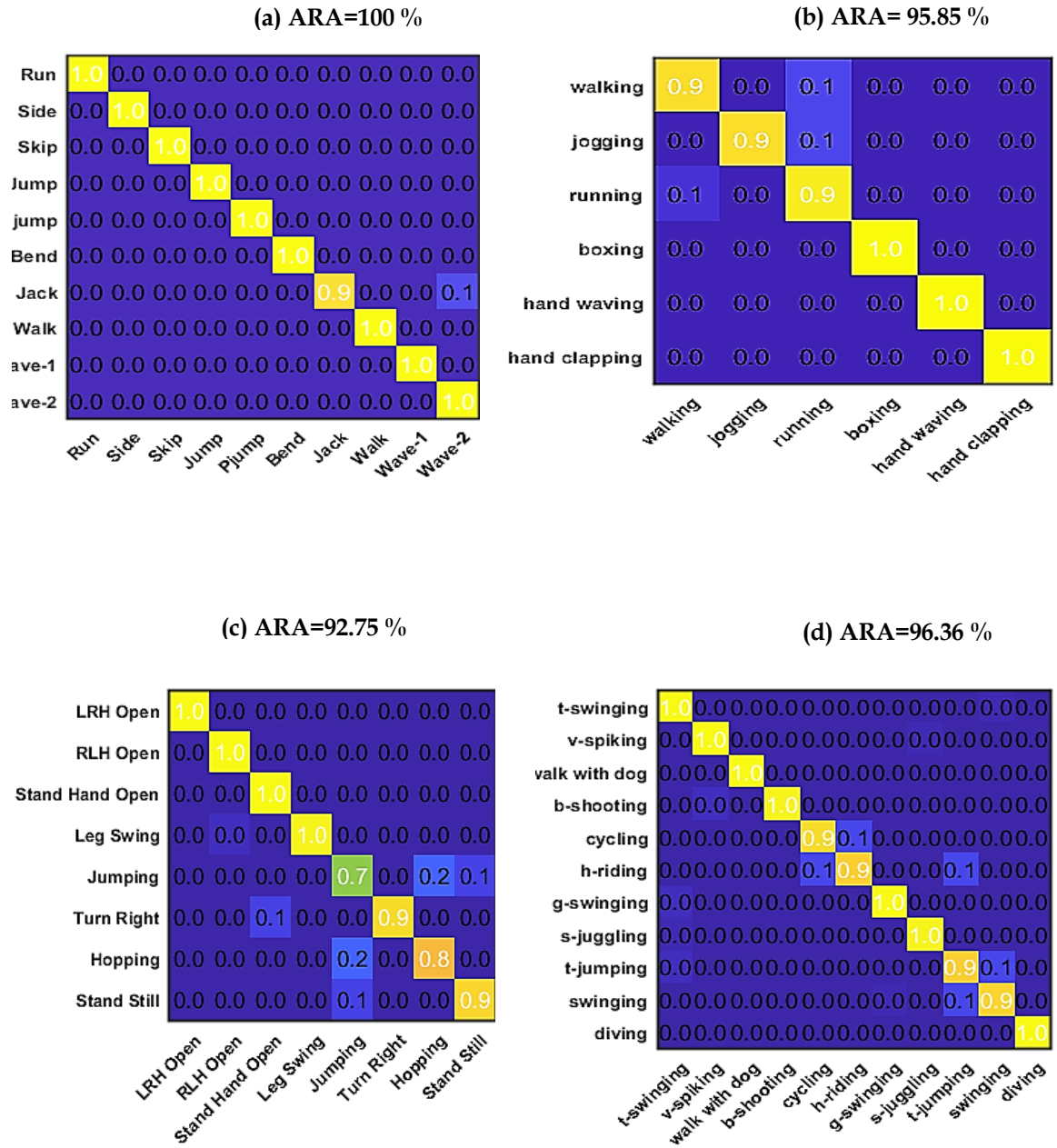


Figure 3.13: Classification result of SVM classifier on (a) Weizmann Action (b) KTH (c) Ballet Movement (d) UCF YouTube action datasets

3.6 Comparison of EMRFs with State-of-the-Art Approaches

The results comparison developed EMRFs with state-of-the-art approaches is illustrated in Tables 3.1, 3.2, 3.3, and 3.4. The comparison is carried through in terms of earlier works, various form of input data to the features extractor, techniques, evaluation protocol, types of classifiers used for classification of action, and highest recognition accuracy. It can be observed from tables that most of the approaches for action recognition in video sequences relied on spatial-temporal features [99] [102] [114] [164] [75] [165] [166]. There are very few approaches using a single image or still image data information [100] [167] [168]. Because it is a challenging task to extract robust features from spatial cue only and in the absence of temporal information. The proposed EMRFs approach with SVM classifier

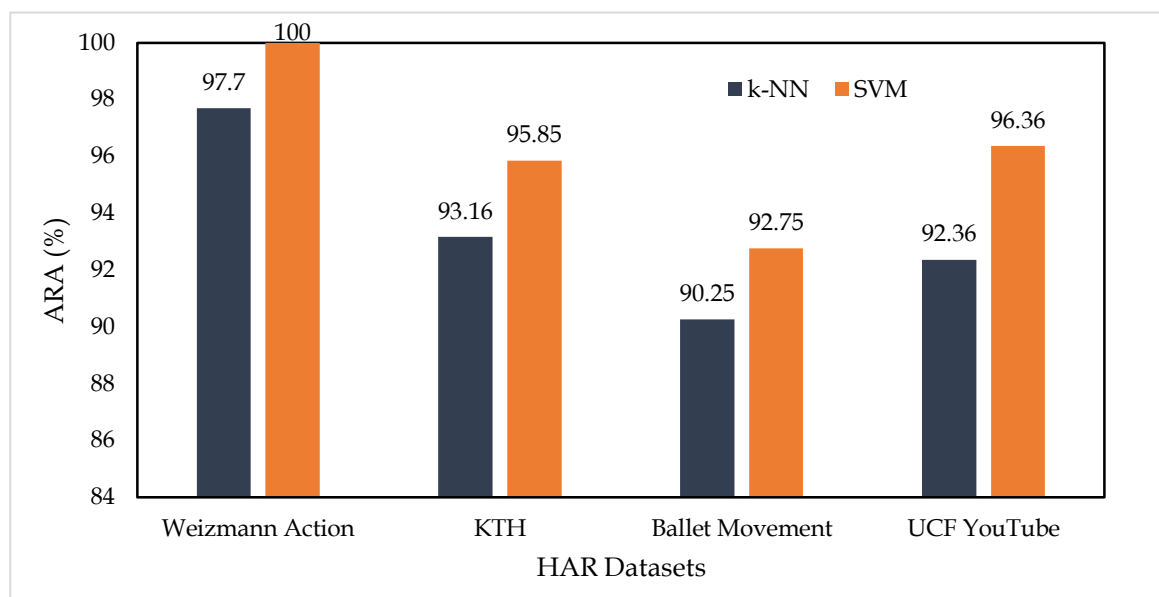


Figure 3.14: ARA Comparison of k-NN and SVM Classifier on HAR datasets

achieved the highest accuracies on the action datasets such as Weizmann, KTH, and UCF YouTube is compared with other state-of-the-arts. The ARA achieved on Weizmann human action dataset is 100%. The main reason for achieving such a high recognition rate is because of simple atomic action performed in control way under non-varying environmental conditions.

The result comparison for the Weizmann and KTH datasets is shown in Tables 3.1 and 3.2, respectively. The comparative analysis with a similar state-of-the-art on KTH dataset is depicted in Table 3.2. For KTH dataset, the action classification ARA equal to 95.83%. The drop of recognition rate in this dataset as compared to Weizmann dataset is due to the more varying illumination and challenging environmental conditions. It is also observed that EMRFs gives a significant amount of increase in recognition accuracy.

The comparison of the various state-of-the-art methods on Ballet dataset is listed in Table 3.3. It is considered a challenging dataset regarding the complexity of human activities performed such as the speed of action and low illumination, but due to enclosed setup conditions, and EMRFs features extraction results for the dataset is practically satisfactory.

Table 3.1: Result comparison with the state-of-the-art on Weizmann Action Dataset

Works	Input	Method	Classifier	Test scheme	ARA (%)
Niebles and Fei [169]	Spatiotemporal	BoFs	SVM	LOOCV	55.00
Thurau [170]	Temporal	HOG	SVM	LOOCV	57.45
Eweiwi et al. [171]	Still Image	NMF	Bayesian	-	55.20
Thurau and Hlavac [99]	Spatiotemporal	HOG	1-NN classifier	LOOCV	74.40
Chaaaraoui et al. [168]	Still Image	Silhouettes	SVM	LOSO	92.80
Baysal and Duygulu [172]	Temporal	GPB, DTW	K-NN	LOOCV	95.10
Guan et al. [101]	Still Image	TNMF	-	CV	91.70
Batchuluun et al. [173]	Temporal	Silhouettes	Fuzzy Logic	-	99.20
Ours	Still Image	EMRFs	SVM	LOOCV	100

Table 3.2: Result comparison with the state-of-the-art on KTH Dataset

Works	Input	Method	Classifier	Test scheme	ARA (%)
Raja et al. [100]	Still Image	HOG	LSVM	-	86.58
Baysal and Duygulu [172]	Temporal	GPB	k-NN	LOOCV	81.30
Saghafi and Rajan [164]	Spatiotemporal	PDE	-	LOOCV	92.60
Han et al. [114]	Spatiotemporal	Deeper spatial ConvNets		Splits	61.11
Zheng et al. [163]	Temporal	Fisher vector	LSVM	Splits	94.58
Ours	Still Image	EMRF	SVM	LOOCV	95.83

Table 3.3: Result comparison with the state-of-the-art on Ballet Dataset

Works	Input	Method	Classifiers	Test Scheme	ARA (%)
Fathi & Mori [160]	Temporal	Optical flow	Adaboost	LOOCV	51.00
Wang and Mori [174]	Temporal	BoWs	S-CTM	LOO	91.30
Guha and Ward [75]	Spatiotemporal	Cuboids+ LMP	RSR	LOO	91.10
Iosifidious et al. [165]	Temporal	BoWs	SVM	LOO	91.10
Zhao et al. [102]	Spatiotemporal	RKHS	K-means	-	79.78
Wang et al. [167]	Still Image	LGLRR	K-means	-	60.87
Vishwakarma et al. [166]	Spatiotemporal	LDA	SVM-NN	LOOCV	94.00
Ours	Still Image	EMRFs	SVM	LOOCV	92.75

The experimental setup used for this dataset is similar to work [166]. The average recognition accuracy achieved is 92.75% which is better than [102] [167] [165] but less than [166]. The main reason for the slightly less accuracy is using of spatial shape feature only in still image of complex motion action dataset.

Table 3.4: Result comparison with the state-of-the-art on UCF YouTube Dataset

Works	Input	Method	Classifiers	Test Scheme	ARA (%)
Liu et al. [2]	Spatiotemporal	Hybrid features	Adaboost	LOOCV	71.20
Cinbis and Sclaroff [175]	Spatiotemporal	MIL	SVM	LOOCV	75.20
Le et al. [176]	Spatiotemporal	Independent sub space analysis	K-Means	-	75.80
Yi and Lin [177]	Spatiotemporal	Spatio-temporal graph	-	LOO	84.63
Wang et al. [9]	Spatiotemporal	Dense trajectories	-	LOOCV	85.40
Shao et al. [178]	Spatiotemporal	Kernel multi-view projection	Naïve Bayes	5-fold CV	87.60
Jung and Hong [179]	Temporal	Bag of Sequence lets	SVM	LOOCV	89.90
Nazir et al. [134]	Spatiotemporal	Bag of Expression(BoE)	KNN	-	96.68
Ours	Still Image	EMRFs	SVM	LOOCV	96.36

Table 3.4 consist of various state-of-the-art approaches on UCF YouTube datasets. It is a more challenging activity dataset because most of the video was recorded in a free environment and complex background conditions. The developed method gave 96.36% recognition accuracy with spatial cues only which outperform over existing spatiotemporal approaches on this complex dataset.

3.7 Significant Outcomes

The above-mentioned study is performed to address the problem of human activity recognition in single still key pose images obtained from video sequences under the challenging conditions such as the absence of temporal information, scale variation, illumination changes, and view variations. After doing this empirical study, we observed the following significance outcomes as:

- It can be noted from Tables 3.1,3.2,3.3, and 3.4 that the proposed approach for action representation showed better results as compared with existing Spatio-temporal approaches using single still images only.
- The proposed feature descriptor shows higher accuracy with two best discriminative classification techniques such as supervised SVM and non-parametric k-nearest neighbour (k-NN). However, the classification accuracy of k-NN classifier somewhat less as compared with SVM classifier because k-NN required higher dimensional data for training the model.
- It is observed that as the number of key poses increase slightly increase the classification accuracy at the cost of increasing complexity of the model. On the other hand, if we increase the number of levels, it has shown effective accuracy, but a higher dimension of the feature vector.

Chapter 4

Learned features based Action Recognition

This chapter introduced the two automatic learned deep frameworks for human activity recognition in RGB and RGB-D(depth) videos. The first deeply coupled ConvNet model based on transfer learning that utilized RGB only frames and dynamic images for the representation of complex actions in videos. On the other hand, our second approach utilized and fused three different modalities RGB, RGB-D(depth) and 3D coordinate information for activity classification for better action recognition and complete utilization of information available from a depth sensor video simultaneously. Further, the classification results of both deep learning approaches are validated on standard depth action datasets and compared with existing state-of-the-art methods.

4.1 Introduction

This chapter work is motivated by the tremendous achievement of deep learning models for computer vision tasks, particularly for human activity recognition. It is gaining more attention due to the numerous applications in real life, for example, smart surveillance system, human-computer interaction, sports action analysis, elderly healthcare etc. Recent days, the acquisition and interface of multimodal data are straightforward due to the invention of low-cost depth devices. Several approaches have been developed based on RGB-D (depth) evidence at the cost of additional equipment's set up and high complexity. Contrarily, the methods that utilize RGB frames only provide inferior performance due to the absence of depth evidence; however, these approaches need less hardware, simple and easy to generalize using only colour cameras. In the first part of this chapter, Section 4.2 de-

scribed the hybrid two-stream ConvNet architecture for human activity recognition by utilized only RGB frames. In the second part of this chapter, Section 4.3 explained the details architecture of proposed deep multimodal network based on bottleneck layer features fusion for action recognition. The details architecture of both the proposed approaches are explained in the following sections:

4.2 A Deeply Coupled ConvNet for Human Activity Recognition using Dynamic and RGB Images

Human action recognition in video sequences motivated and extended by the improved image recognition approaches. Most of the video action recognition approaches based on shallow higher dimensional spatiotemporal features extraction from stacked raw video frames. Human action motion can be disintegrated into spatial and temporal features in a video. The spatial features contain the appearance information about an object in each sequence of a given video. On the other hand, temporal features represented in the form object moving across the video sequences. The proposed human action recognition model accordingly divided into two streams model as depicted in Fig.4.1.

In this work, a hybrid two-stream deep architecture based on two different data streams (spatial and temporal) is developed that is then fused by late fusion techniques to recognize the activities. The spatial features are extracted from RGB frames fed at regular interval to recognize action while the temporal stream is trained using the dynamic motion images that captured the full temporal dynamics of a video. Both the streams are trained using pre-trained Inception-v3 deep architecture [180]. The first spatial stream of the network is trained end-to-end learning through a pre-trained ConvNet and followed by the Bi-LSTM network for additional sequential information representation. The second temporal stream of the network is fine-tuned on the pre-trained ConvNet. These streams of networks are connected parallelly, as shown in Fig.4.1. Further, the scores obtained

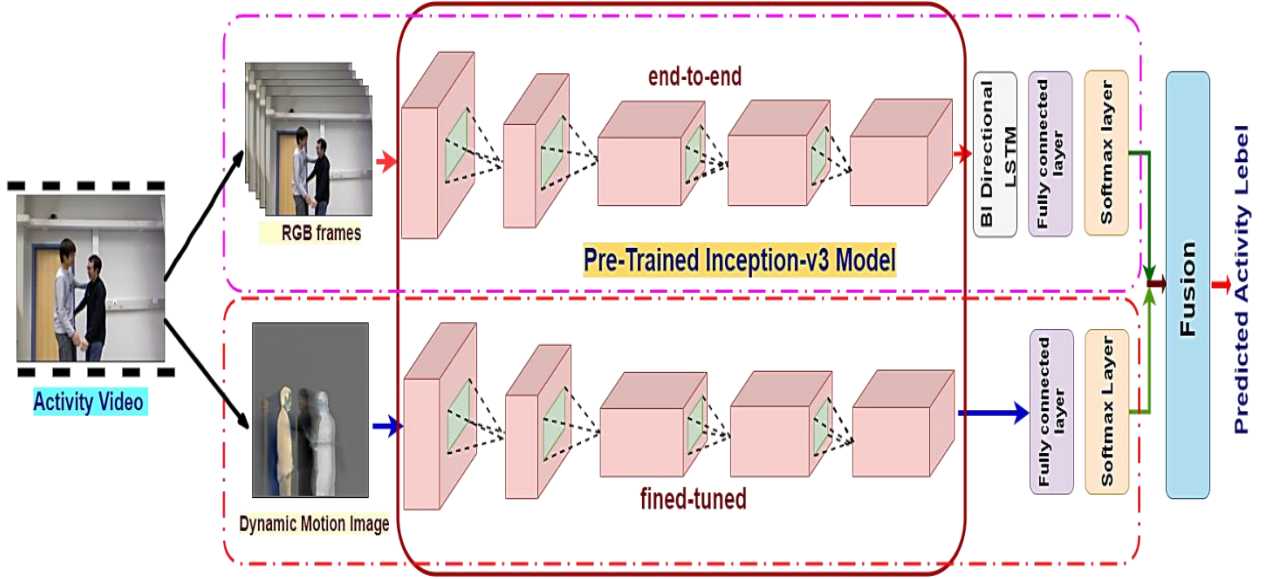


Figure 4.1: Schema of Deeply Coupled ConvNet for Human Activity Recognition using Dynamic and RGB Images

by these two streams of the network are fused with late fusion techniques at the decision level after the softmax layer to enhance the classification accuracy of the proposed model.

4.2.1 Features Extraction with Pre-Trained Inception-v3 Architecture

In this section, we briefly discussed the underlying architecture of deep pre-trained model that is utilized for fine-tuning and feature extraction from the input video. Fig. 4.2 depicted the deep Inception-v3 architecture consisted of one input block, three blocks of Inception Module A, B and C, two blocks of grid size reduction, one Auxiliary classifier block and one output block. This deep network model enriched with some advanced features such as RMSProp optimizer, Batch Normalization, Label Smoothing to reduced overfitting and add loss function as compared with the previous version of inception architectures. This network consists of 42 deep layers that accept input data $299 \times 299 \times 3$ of spatial dimension at input block. The Inception Module A is used for smaller factorization convolutional (Conv) and converted a 5×5 (Conv) filter into two 3×3 Conv filters which re-

sulted in reducing parameters between the layers without decreasing the efficiency of the network. The Inception Module B utilized for asymmetric spatial factorization convolutions which converts a 3×3 Conv filter into 1×3 Conv followed by 3×1 Conv filters. It can be observed that a $n \times n$ Conv can be represented by $1 \times n$ Conv followed by a $n \times 1$ Conv save the computational cost and reduced the overall parameters. The Inception Module C is introduced for stimulating the high dimensional representations similarly works as module B in this network. The grid size-reduction block is used for downsizing the feature map such as in deep AlexNet or VGGNet models. The main difference between traditional models and inception-v3 is that a $m \times m$ grid with n filters is divided into $m/2 \times m/2$ grid with $2n$ filters. Thus the overall computational cost is decreased by using convolutional operation followed by pooling operation. This network contains one auxiliary classifier on the top of the last 17×17 layer which is used as a Regularizer for enhancing the convergence of the deep network.

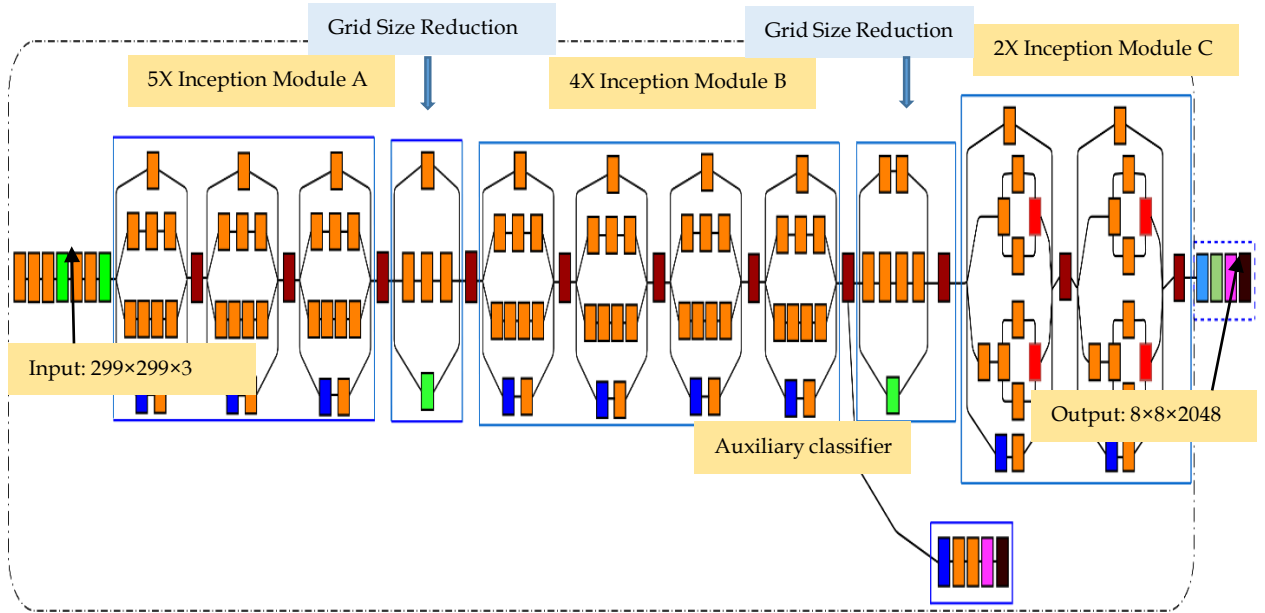


Figure 4.2: Block diagram of Inception-v3 improved Deep Architecture

In this work, the deep pre-trained network is trained in an end-to-end manner followed by the Bi-LSTM architecture which represents one of the action descriptors for the two-stream HAR model. In the following sub-section, we briefly discussed the underlying architecture of LSTM [181] and Bi-LSTM that is utilized for additional sequential features extractions.

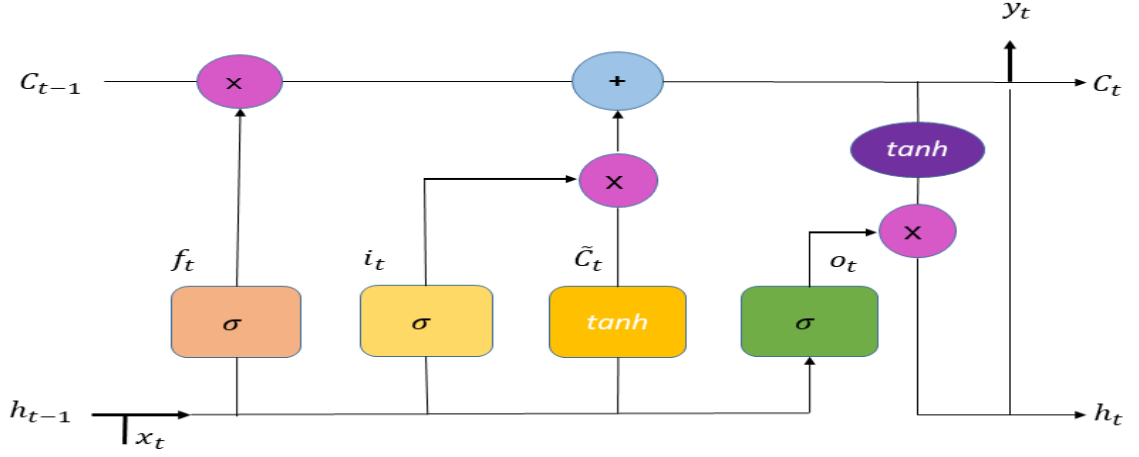


Figure 4.3: Basic LSTM Architecture

4.2.2 The Bi-Directional LSTM (Bi-LSTM)

It is observed that the traditional feedforward network and convolutional neural network are inefficient to deal with sequential data such as video analysis. Such networks accept fixed-length input videos. To overcome the problem of fixed size, input padding is used but the performance of such approaches are not comparable with RNNs [182] and LSTM network [181]. The RNN deep network has compatibility with the help of the chain and loop structure to deal with sequential data. However, RNNs are inefficient as compared to LSTM for longer duration sequential input because of the vanishing gradient in the backpropagation process. Due to the problem of vanishing gradient, the LSTM architecture was proposed, which utilized two parallel RNNs for enhancing the long term dependencies and training for bigger input networks. The gated structure of the LSTM network helped the many sequential learning problems in

a very effective way.

Fig. 4.3 depicted the basic architecture of the LSTM network. It can be observed from that LSTM network has a similar chain-like structure like RNN. It consisted of four neural layers, while RNN is having only one neural sigmoid layer (\tanh). A horizontal line at the top of Fig. 4.3 is called the 'cell state' of LSTM which works like a conveyor belt. This cell state has the capabilities to remove and add information. The LSTM is consists of three gates: input (i_t), forget and output (o_t) gates to control and protect the cell states. These each gates having sigmoid neural layer and capable of process the information through pointwise multiplication units.

In the beginning, the 'forget gate (f_t)' decides what information between (h_{t-1}, x_t) is passed through the cell states and defined as Eq.(4.1).

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (4.1)$$

where, the activation function is denoted by the $\sigma(\cdot)$ and x_t, h_{t-1}, W_f, b_f represent the input, output at previous LSTM block, weight, and bias at the forget gate layer respectively at time t . On the next step, the input gate (i_t), decides what new information is to store in the cell state. The input gate equation is given as Eq.(4.2).

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (4.2)$$

where, W_i , and b_i represent the weight and bias at the input gate layer, respectively.

For the update, a new state in the cell state, a sigmoid function creates a new vector \hat{C}_t as defined by Eq.(4.3). Later on, Eq.(4.2) and Eq.(4.3) are combined for updating a new state.

$$\hat{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (4.3)$$

In order to update the old cell (C_{t-1}) into new (C_t), the old state is multiplied with forget gate and add other parameters as shown in Eq.(4.4).

$$C_t = f_t * C_{t-1} + C_{t-1} + i_t * \hat{C}_t \quad (4.4)$$

Finally, the output o_t is given the cell state based on the output of the sigmoid of output gates and defined by:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (4.5)$$

$$h_t = o_t * \tanh(C_t) \quad (4.6)$$

where W_o , and b_o represent the weight and bias at the output layer, respectively.

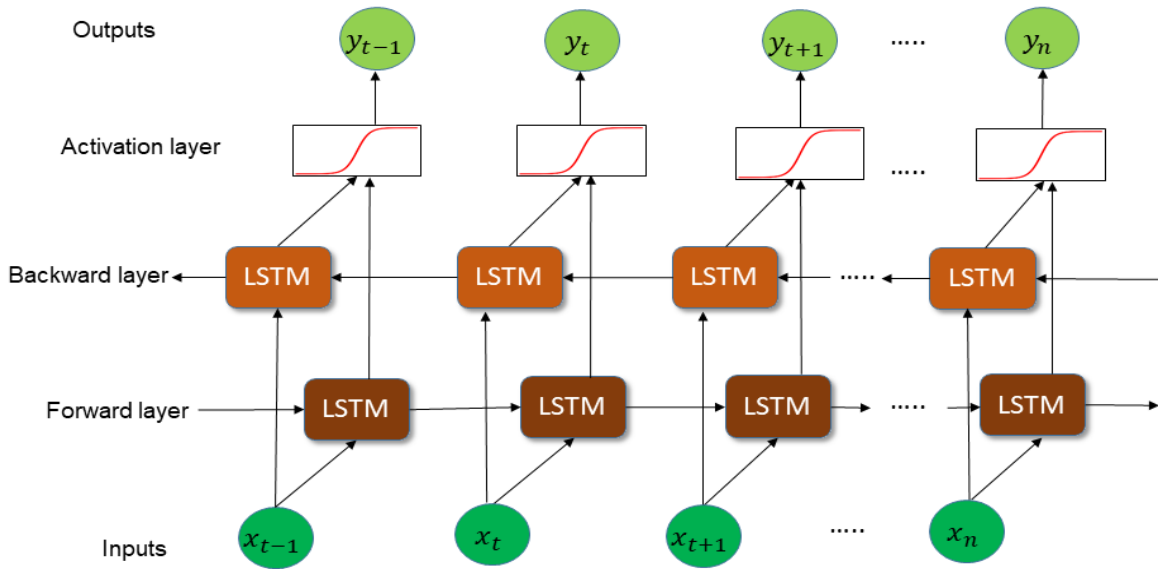


Figure 4.4: The Bi-directional LSTM Architecture

It is noticed that unidirectional LSTM network stores the information data from the past direction only. On the other hand, a Bi-LSTM network preserves the information data both direction from past to future as well as the future to past. Further, the Bi-LSTM network performed much better than LSTM for sequential data applications. The basic architecture of a Bi-LSTM network depicted in Fig. 4.4. The output sequence in the forward layer \vec{h} is computed for given inputs in positive sequence from $x_{t-1} \dots x_n$. The backward output sequence \tilde{h} is computed for reversed inputs sequences. The final layer output is evaluated with the help of

Equations (4.1) - (4.6) for the backward and forward layer outputs. The bi-directional LSTM layer yields an output vector as: $Y_T = y_{t-1}, y_t, \dots, y_{t+n}$ in which each $y_t = \sigma(\vec{h}, \overleftarrow{h})$, where σ is called a summation function or concatenating function.

It is observed that fine-tune the all sequential frames related to action labels is not a good idea for learning good discriminative features. Therefore, in this work, RGB video is sampled out ten frames at a fixed interval of to the pre-trained CNN architecture and trained excluding the fully connected layer. The upper half of the proposed ConvNet is end-to-end trained with only the last five layers. For each frame, a features vector of dimension $10 \times 4 \times 4 \times 2048$ is obtained from CNN architecture. Further, these extracted features are given to average pooling layer and obtained a vector of dimension $10 \times 1 \times 2048$ for each layer. Finally, the feature vectors acquired from the pre-trained model is further fed to the Bi-LSTM layer for better feature extractions. It helps for better generalization in terms of learning features from frames. In the following subsection, the concept of dynamic motion images and fine-tuning with CNN for temporal extraction features are explained in detail.

4.2.3 Dynamic motion image (DMI) from video sequences

In this section, we emphasized understanding the long term temporal dynamics in terms of a single dynamic image. Later on, these images are fed as an input for fine-tuned inception-v3 architecture for temporal features extraction. It is a paramount task to understand the content of videos precisely on a large scale because of videos are consisting of a sequence of still images. Therefore, the summarization of the whole video sequence into a single still dynamic image using standard CNN architecture introduced in work [183]. A rank pooling mechanism is adopted that utilized the work of Fernando et al. [130] for obtaining a dynamic image from whole video sequences. This technique encrypts the temporal variation cues of the video sequences into one single image. A rank pooling function directly applied to raw RGB frames to produces a single dynamic image for each activity video.

The idea behind the creation of dynamic images depends on the ranking function [130] that ranks its each video frames $(I_1, I_2, I_3, \dots, I_T)$ according to the time axis. Let a feature vector represented by $\varphi(I_t) \in \mathbb{R}^{\mathbf{d}}$, $t \in [1, T]$, where $\varphi(\cdot)$ is showed the rank for each frame I_t at time instance t . A score function $\mathcal{S}(t | \mathbf{d}) = \langle \mathbf{d}, \mathcal{V}_t \rangle$, is associated with ranking function, where $\mathbf{d} \in \mathbb{R}^{\mathbf{d}}$ is a vector of parameters and $\mathcal{V}_t = \frac{1}{t} \sum_{\tau=1}^t \varphi(I_\tau)$ is the time average of these features at this instant. According to the RankSVM [184], the learning of vector \mathbf{d} is modelled as a convex optimization problem and given by Eq. (4.7),

$$\mathbf{d}^* = \rho(I_1, I_2, I_3, \dots, I_T; \varphi) = \underbrace{\operatorname{argmin}_{\mathbf{d}}}_{\mathbf{d}} E(\mathbf{d}) \quad (4.7)$$

$$E(\mathbf{d}) = \frac{\delta}{2} \|\mathbf{d}\|^2 + \frac{2}{T(T-1)} \times \sum_{q>t} \max\{0, 1 - \mathcal{S}(q|\mathbf{d}) - \mathcal{S}(t|\mathbf{d})\} \quad (4.8)$$

where $\rho(I_1, I_2, I_3, \dots, I_T; \varphi)$ maps a sequence of T number of video frames into a single vector \mathbf{d}^* called a rank pooling function that aggregates information from all video frames. The first term in Eq. (4.8), $\frac{\delta}{2} \|\mathbf{d}\|^2$ is a quadratic regularized function used for SVM. The second term related to soft counting loss that calculates how many pairs are not correctly ranked for $q > t$ by the ranking function. The score function is calculated for video frames based on ranking function and a pair of the frame is chosen for which score having unit margin i.e. $\mathcal{S}(q|\mathbf{d}) > \mathcal{S}(t|\mathbf{d}) + 1$.

The dynamic motion image obtained using approximate rank pooling function [183] is fifty times faster than rank pooling function for similar performance. The approximate rank pooling mechanism is worked on a gradient optimization algorithm and derived using Eq. (4.9) to (4.12) as follows:

for $\mathbf{d} = \mathbf{0}$, $\mathbf{d}^* = \vec{0} - \eta \nabla E(\mathbf{d})|_{\mathbf{d}=\vec{0}} \propto -\nabla E(\mathbf{d})|_{\mathbf{d}=\vec{0}}$ for all $\eta > 0$ where,

$$\begin{aligned} \nabla E(\vec{0}) &\propto \sum_{q>t} \max\{0, 1 - \mathcal{S}(q|\mathbf{d}) - \mathcal{S}(t|\mathbf{d})\} \Big|_{\mathbf{d}=\vec{0}} \\ &= \sum_{q>t} \nabla \langle \mathbf{d}, \mathcal{V}_t - \mathcal{V}_q \rangle = \sum_{q>t} \mathcal{V}_t - \mathcal{V}_q \end{aligned} \quad (4.9)$$

Further, the function \mathbf{d}^* used as video descriptor because it aggregates information from all stacks frame and defined as:

$$\mathbf{d}^* \propto \sum_{q>t} \mathcal{V}_t - \mathcal{V}_q = \sum_{q>t} \left[\frac{1}{q} \sum_{i=1}^q \varphi_i - \frac{1}{t} \sum_{j=1}^t \varphi_j \right] = \sum_{t=1}^T \Omega_t \varphi_t \quad (4.10)$$

where, the Ω_t is denoted as:

$$\Omega_t = 2(T - t + 1) - (T + 1)(H_T - H_{t-1}) \quad (4.11)$$

where, $H_t = \sum_{i=1}^T \frac{1}{i}$ is the i^{th} Harmonic number and $H_0 = 0$.

Therefore, the updated approximate rank pooling function is defined as weighted sum of adjacent video frames as:

$$\hat{\rho}(I_1, I_2, I_3, \dots, I_T; \varphi) = \sum_{t=1}^T \Omega_t \varphi_t \quad (4.12)$$

The weights parameter $\Omega_t \in (1, T)$ is calculated for fixed video length accordingly, as shown in the Fig.4.5. This demonstration showed that Ω_t depends on the consecutive video frames belongs to frames $(1, T)$. The weight parameter Ω_T is the combination of the sum of all weight parameters obtained from each frame $\frac{2 \cdot i - T - 1}{i}$, where $i \in (t, T)$ by using Eq.(4.10). It showed that DMIs computation is limited only to by pre-multiplying function Ω_t for all video frames. Therefore, approximate rank pooling function does not require to compute the intermediate average features vector $\mathcal{V}_t = \frac{1}{t} \sum_{\tau=1}^t \varphi(I_\tau)$. Instead, it can be directly calculated

by using individual frame feature $\varphi(I_t)$ and $\Omega_t = 2t - T - 1$, which is a linear function of time t .

		I_1	\oplus	I_2	\oplus	I_3	$\dots\dots\dots$	I_{T-1}	\oplus	I_T
$\Omega_1 \rightarrow$		$\frac{2*1-T-1}{1}$		$\frac{2*2-T-1}{2}$		$\frac{2*3-T-1}{3}$	$\dots\dots\dots$	$\frac{2*(T-1)-T-1}{T-1}$		$\frac{2*(T)-T-1}{T}$
$\Omega_2 \rightarrow$				$\frac{2*2-T-1}{2}$		$\frac{2*3-T-1}{3}$	$\dots\dots\dots$	$\frac{2*(T-1)-T-1}{(T-1)}$		$\frac{2*(T)-T-1}{T}$
$\Omega_3 \rightarrow$						$\frac{2*3-T-1}{3}$	$\dots\dots\dots$	$\frac{2*(T-1)-T-1}{(T-1)}$		$\frac{2*(T)-T-1}{T}$
\cdot										
\cdot										
\cdot										
$\Omega_{T-1} \rightarrow$								$\frac{2*(T-1)-T-1}{(T-1)}$		$\frac{2*(T)-T-1}{T}$
$\Omega_T \rightarrow$										$\frac{2*(T)-T-1}{T}$

Figure 4.5: Shown the process of calculation of parameter Ω_T for finite length video sequences T . The bold part shows the dependency of parameter Ω_T on consecutive video frames $\in (1, T)$.

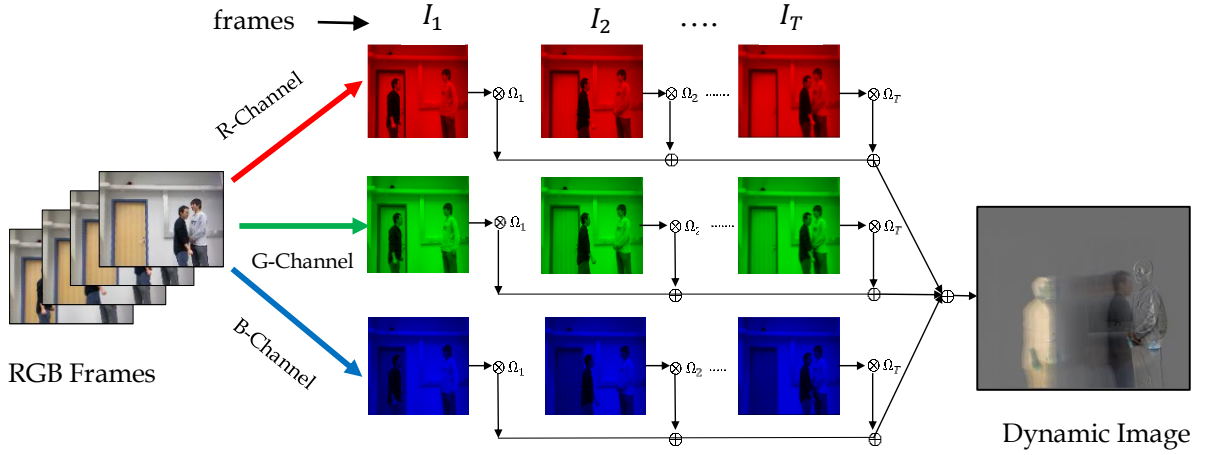


Figure 4.6: Shown the formation of Dynamic Motion Image using Approximate Rank Pooling Mechanism from Red, Green and Blue channel of each video frame

The example of dynamic image formation using the approximate rank pooling technique from 'approaching' activity video is illustrated in Fig.4.6. It can be noted that each video frames is multiplied with the corresponding weight, i.e. frame I_1

is multiplied by Ω_1 for each channel individually. The final obtained DMI is the weighted sum of each Red, Green, and Blue channels of each frame and of the same size as the frames. It is observed that the temporal action modelling pattern can be easily seen from a dynamic image when a person is approaching another person irrespective to background and illumination conditions. The so obtained DMIs are fed to pre-trained ConvNet to extract the temporal features to recognize the action in a video.

4.2.4 Late Fusion

The late fusion techniques for two-stream networks are categorized as Sum Fusion, Maximum fusion and Concatenation Fusion [117]. The main objective of fusing the two-stream network is that features extracted from the same pixel's location through the different channel are combining for better prediction. For example, differentiating between the activities of 'talking on mobile phone' and 'drinking water through glass'. To discriminate these activities, a hand movement pattern can be easily recognized by the temporal network at some spatial location while then the spatial network can identify the location of the head and their combination help to increase the overall prediction accuracy. In the proposed work, these techniques are defined using the scores of a decision level of Inception-Bi-LSTM stream (a) and DMI stream (b).

A general function f for fusing the two feature maps a and b at a given time t is denoted by Eq.(4.13).

$$f: x_t^a, x_t^b \rightarrow y_t \quad (4.13)$$

where, $x_t^a \in \mathbb{R}^{H^a \times W^a \times D^a}$, $x_t^b \in \mathbb{R}^{H^b \times W^b \times D^b}$ are two different features maps respectively.

The output feature map is denoted as: $y_t \in \mathbb{R}^{H \times W \times D}$, where W , H and D are represented width, height and number of the channel of respective feature maps. For simplicity, we assume that $W^a = W^b = W$, $H^a = H^b = H$, $D^a = D^b = D$. The late

fused score obtained by these approaches are denoted as y^{sum} (Sum), y^{max} (Maximum), and y^{cat} (Concatenation).

Sum Fusion: It calculates the sum $y^{sum} = f^{sum}(x^a, x^b)$ of two features maps in the feature channels d , at the same spatial location (i, j) and expressed using in Eq.4.14.

$$y_{i,j,d}^{sum} = x_{i,j,d}^a + x_{i,j,d}^b \quad (4.14)$$

where, $1 \leq i \leq H$ $1 \leq j \leq W$ $1 \leq d \leq D$ and $x^a, x^b, y \in \mathbb{R}^{H \times W \times D}$. The sum fusion scores y^{sum} show a random correlation between the network layers.

Maximum fusion: In this technique, $y^{max} = f^{max}(x^a, x^b)$ the maximum score is selected between the two feature maps. It defined using Eq. 4.15.

$$y_{i,j,d}^{max} = \max\{x_{i,j,d}^a + x_{i,j,d}^b\} \quad (4.15)$$

where all the parameter representations are similar as in Eq. (4.14). It increases accuracy as it incorporates best the predictions of both the models and this approach is used in our model.

Concatenation Fusion: In this fusion method, $y^{cat} = f^{cat}(x^a, x^b)$ features extracted from both streams are stacked across the feature channel d as:

$$y_{i,j,2d}^{cat} = x_{i,j,d}^a \quad y_{i,j,2d-1}^{cat} = x_{i,j,d}^b \quad (4.16)$$

here $y_t \in \mathbb{R}^{H \times W \times 2D}$, the concatenation fusion does not show any correlation between the two feature maps.

4.2.5 Implementation Details

In the following section, model descriptions, training, results and comparison with similar state-of-the-art approaches on four standard RGB-D datasets are discussed in details. In the proposed ConvNet the deep network is trained on one NVIDIA GTX 2GB graphic with 8GB RAM, GPU machine. The Keras API deep learning library is used for the implementation of two streams of convolutional networks.

To evaluate the performance of our two-stream network, we use the four standard datasets one is focused on a dyadic activity as SBU Interaction [185], single and human-object interaction activity datasets as MIVIA Action [83], MSR Daily Activity [186] and MSR Action Pairs [187]. The detailed description of these datasets such as a number of actions, actors, and challenges are explained in the following subsection.

4.2.5.1 The SBU Interaction Dataset

Yun et al. [185] introduced this human interaction activity dataset. It is recorded with three different modalities RGB, depth frames, and 3D coordinates with the help of Kinect Sensor. It consists of 7 subjects performing 8 human-human interaction activities: departing (S1), approaching (S2), hugging (S3), pushing (S4), kicking (S5), punching (S6), exchanging object (S7), and shaking hands (S8). The sample images from this dataset are depicted in Fig.4.7.

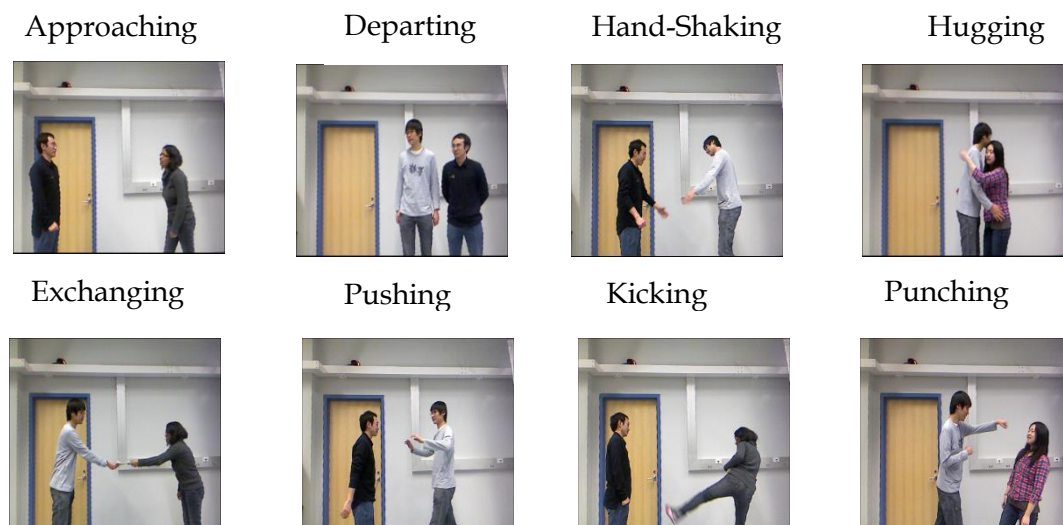


Figure 4.7: Sample RGB frames from SBU Interaction dataset

4.2.5.2 MIVIA Action

Carletti et al. [83] introduced this dataset intending to recognize human-object interaction in an indoor lab environment. This dataset is recorded with two different modalities RGB and depth frames with the of Kinect sensors. It contains 14 actors (7 females and 7 males) performing 7 activities such as: “drinking” (M1), “sleep-

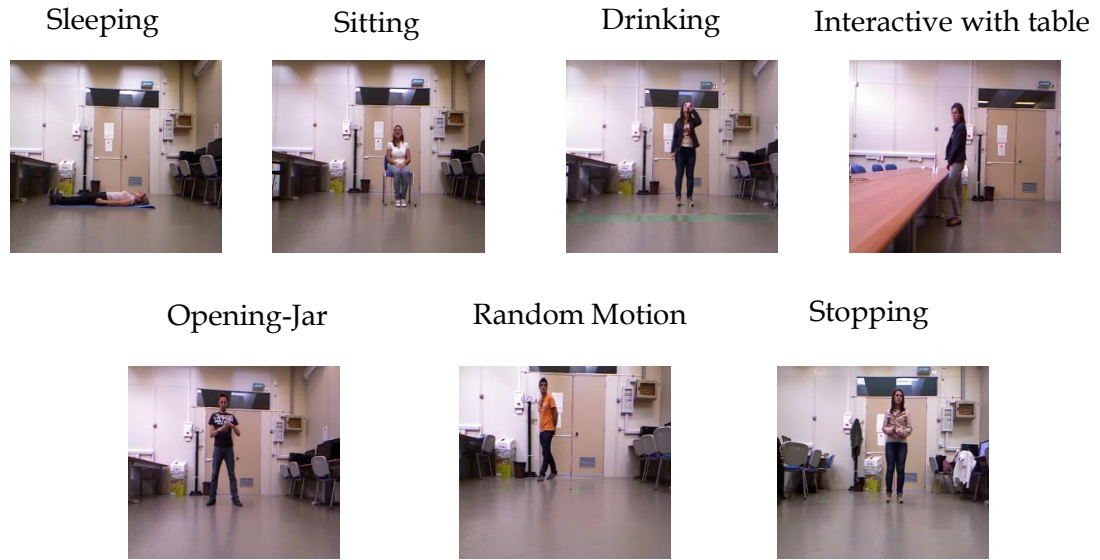


Figure 4.8: Sample RGB frames from MIVIA Action dataset

ing” (M2), “opening a jar” (M3), “sitting” (M4), “interacting with a table” (M5), “stopping” (M6), and “random motion” (M7). The sample images from this dataset are depicted in Fig.4.8. testing and 5 for training purposes. The sample images from this dataset is depicted in Fig.4.9.

4.2.5.3 MSR Action Pairs 3D

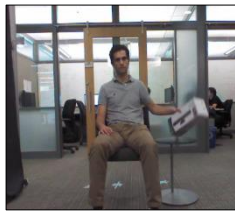
Oreifej and Liu [187] introduced this dataset consists of pairs of action videos. It is a challenging dataset because similar activities have the same shape and motion cues such as ‘Put down’ and ‘Put up’. Ten subjects are performing 6 actions pairs: “Pushing and Pulling a chair” (MA1), “Putting and Taking off a backpack” (MA2), “Sticking and Removing a poster” (MA3), “Wearing and Taking off a hat” (MA4),

“Lifting and Placing a box” (MA5), “Picking up and Putting down a box” (MA6). Each action was repeated three times by subjects in which 5 subjects are used for

4.2.5.4 MSR DAILY Activity 3D Dataset

Wang et al. [186] proposed this dataset intending to recognize the daily use of human activity in an indoor room environment. It is recorded with two different modalities: RGB video and depth frames by a Kinect sensor. It contains 10 subjects doing 16 different daily activities: tossing paper (MD1), playing game(MD2), stand up(MD3), playing guitar(MD4), walking(MD5), using laptop(MD6), cheer up(MD7), using vacuum cleaner(MD8), calling on cell phone (MD9), sit still (MD10), drinking (MD11), lay down on sofa(MD12), reading book(MD13), writing on a paper(MD14), sitting down(MD15), and eating(MD16). All such activities are repeated twice for the sitting and standing position. The sample images from this dataset are shown in Fig. 4.10.

Pick up and Put down



Lift and Place



Push and Pull



Put and Take off



Stick and Remove



Wear and Take



Figure 4.9: Sample RGB frames from MSR Action Pair dataset

Call cell phone



Cheer up



Drink



Write on Paper



Use laptop



Use vacuum cleaner



Walk



Sit Still



Eat



Lie down



Play game



Stand up



Play guitar



Read-book



Sit down



Toss paper



Figure 4.15: Sample RGB frames from MSR Daily Activity dataset

4.2.6 Model Parameter Description and Training Settings

The proposed hybrid ConvNet is trained for two different data streams, i.e. RGB frames and dynamic image independently. The overfitting problem that occurs due to smaller training datasets in LSTM is compensated with the implementation of L2 regularization and dropout mechanism. The CNN-LSTM stream extracted the features from RGB frames that are fed with a batch size of 8 videos. The RGB stream is trained in an end to end fashion up to 150 epochs to refine the features of the pre-trained CNN. Initially, the learning rate of 10^{-4} , and a momentum constant equal to 0.9 is used for training the SGD optimizer. A recurrent dropout of 0.6 is used and added with each Bi-LSTM layer. In the SGD optimizer, the Minimum Square Error (MSE) loss function is selected for calculated the loss during training and test process. Alternatively, the dynamic motion images are fine-tuned with last fully connected layers on the pre-trained CNN model. This DMIs stream is trained on CNN with a batch size of 8 videos up to 150 epochs in Adam Optimizer [188].

The initial learning rate and various parameters in Adam optimizer are used as: 10^{-4} , $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$ respectively, and 0.8 of dropout added after every fully connected layer. We used Adam optimizer for DMIs because it is easy to implement, computationally efficient, requires less memory as compared with other optimizers. It requires less tuning and suitable for non-static patterns objective. Fig. 4.11 illustrated the training and test loss curves using MSE loss function for all four datasets. It is observed from the loss curves that our model attained a lower value of square error near 147 epochs.

All the datasets used in the evaluation of the algorithm is multiclass and therefore, we adopted a multiclass classification cross-validation scheme. The leave one out 5-fold cross-validation multiclass classification scheme is applied for

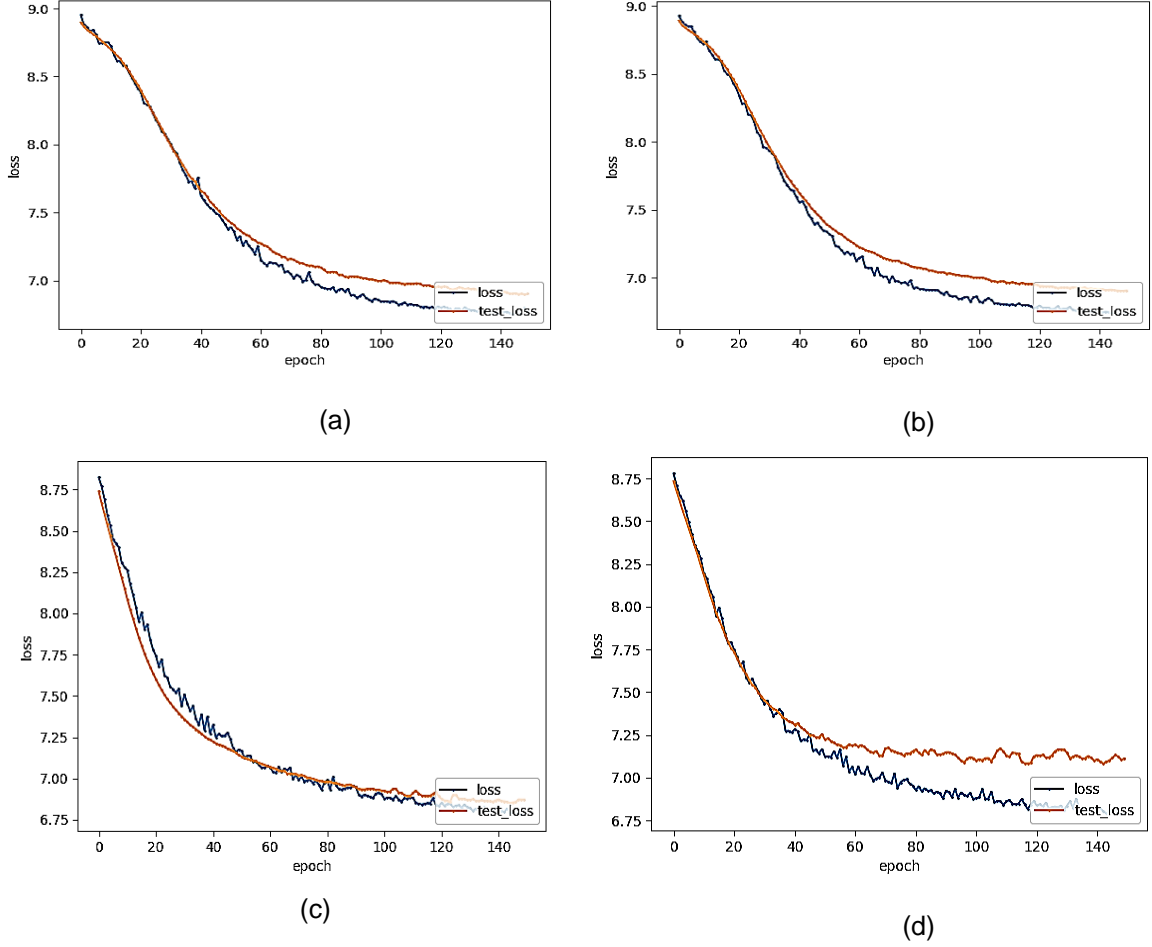


Figure 4.11: Shown the training minimum squared (MSE) loss and test loss for activity datasets: a) SBU Interaction b) Mivia Action c) MSR Action Pairs d) MSR Daily Activity

SBU dataset, where the number of test samples predicted as true class samples are defined as true positives and the test samples predicted as any of other negative classes of the action datasets is considered as false negatives or true negatives. For MIVIA Action dataset, leave-one-subject-out (LOSO)-CV is applied. There is a total of 14 actors performing the 7 human activities. In this evaluation protocol, 13 actors are used for training and the remaining one actor is for testing. This process is repeating 14 times, always leaving another actor's data for testing. The evaluation method adopted for MSR Daily Activity and MSR Action pairs datasets in which half of the subject used for training and half of the subject for testing. The scores generated by both the softmax layers are fused using some late fusion techniques for prediction the final label activity.

4.2.7 Results Analysis and Comparisons

The proposed ConvNet model is tested on four datasets. The evaluation performance is measured in terms of Average Recognition Accuracy (ARA) per class for many classes (\mathcal{C}_i), which is calculated as in Eq.(4.17),

$$ARA = \frac{1}{k} \sum_{i=1}^k \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i} \quad (4.17)$$

where, tp_i true positive of (\mathcal{C}_i), tn_i true negative of (\mathcal{C}_i), fp_i false positive of (\mathcal{C}_i), fn_i false negative of (\mathcal{C}_i).

4.2.8 The Mann-Whitney U Test (Wilcoxon Rank Sum Test)

In order to have a better understanding of classification accuracy, a Wilcoxon Rank Sum Test is used to analyse the result. It is a non-parametric test [189] which is used to test whether two samples are similar in a distribution or not. The null hypothesis that comes from the same observation samples (i.e. have the same median) or alternatively, whether observation in one sample tends to be greater than observed in the other samples. In the proposed work, we have tested the accuracies of both streams i.e. RGB frames and DMI using non-parametric test. Because both streams extracted the features from the same input video sequences independently. The null and alternative hypotheses for the test are stated:

H_0 = the two independent samples accuracy are the same verses,

H_1 = the two independent samples' accuracy is not the same.

This non-parametric test is conducted as a two-tailed test and observed that populations are not the same as opposed to specifying directionality. The test statistic for Wilcoxon Rank Sum Test is represented as U and is chosen from a minimum of U_{RGB} and U_{DMI} given by the following equations.

$$U_{RGB} = n_1 n_2 - \frac{n_1(n_1 + 1)}{2} - \sum R_1 \quad (4.18)$$

$$U_{DMI} = n_1 n_2 - \frac{n_2(n_2 + 1)}{2} - \sum R_2 \quad (4.19)$$

where $\sum R_1$, and $\sum R_2$ are the sum of the ranks for RGB frame samples and DMI samples respectively, n_1 is the total samples of RGB frame accuracy and n_2 is the total samples of DMI accuracy as illustrated in Tables 4.1,4.2,4.3 and 4.4.

Table 4.1: Activity wise Results of RGB Frames and DMI streams on SBU Interaction Dataset

Activity		S1	S2	S3	S4	S5	S6	S7	S8
Accuracy (%)	RGB	82	84	72	71	72	73	76	83
	DMI	98	86	95	96	92	94	82	88

Each time we tested and observed the value of U (U_{Stat}) whether it supports the null or alternative hypothesis like parametric testing. Further, we have determined the critical value ($U_{Cri(0.05 \text{ or } 0.01)}$) and compared with a minimum value of U_{Stat} .

If the critical value is higher, then we reject the null hypothesis H_0 and if the U_{Stat} value is higher than critical value we reject the alternate hypothesis H_1 i.e. $U_{Stat} = \min(U_{RGB}, U_{DMI})$; and if $U_{Stat} < U_{Cri(0.05 \text{ or } 0.01)}$, then H_0 (Rejected), and (Accepted).

Table 4.2: Activity wise Results of RGB Frames and DMI streams on MIVIA Action Dataset

Activity		M1	M2	M3	M4	M5	M6	M7
Accuracy (%)	RGB	82	74	92	85	76	83	79
	DMI	95	94	96	89	94	97	84

Table 4.3: Activity wise Results of RGB Frames and DMI streams on MSR Action Pairs Dataset

Activity		MA1	MA2	MA3	MA4	MA5	MA6
Accuracy (%)	RGB	91	88	90	87	84	86
	DMI	97	96	98	95	87	96

Table 4.4: Activity wise Results of RGB Frames and DMI streams on MSR Daily Activity Dataset

Activity		MD1	MD2	MD3	MD4	MD5	MD6	MD7	MD8
Accuracy (%)	RGB	81	67	82	80	71	66	82	84
	DMI	88	72	90	87	79	72	94	90
Activity		MD9	MD10	MD11	MD12	MD13	MD14	MD15	MD16
Accuracy (%)	RGB	72	73	78	84	80	67	80	79
	DMI	80	81	86	91	89	72	88	84

It can be observed from Table 4.5 that for each dataset the null hypothesis H_0 is rejected and alternate hypothesis H_1 is accepted because in the value of $U_{Stat} < U_{Cri(0.05 \text{ or } 0.01)}$. Therefore, our observation on the samples from independent input streams such as RGB and DMI is not accepted in the same way according to a two-tailed analysis of Wilcoxon rank-sum hypothesis.

Table 4.5: Wilcoxon Rank Sum Test results on four human activity datasets

Dataset	Samples	U_{RGB}	U_{DMI}	U_{Stat}	$U_{Critical}$	
	$n_1(RGB) = n_2(DMI)$				$\alpha = 0.05$	$\alpha = 0.01$
SBU Interaction	8	61.50	2.5	2.5	13	7
Mivia Action	7	46	3	3	8	4
MSR Action Pairs	6	32.5	3.5	3.5	5	2
MSR Daily Activity	16	201	55	55	75	60

The result obtained by various late fusion techniques is compared to four datasets in Table 4.6. It can be observed that max fusion and sum fusion techniques give high scores as compared to average and concatenation fusion technique. The

max fusion yields highest recognition accuracy at softmax layers because it selects the maximum probability from both of the softmax prediction scores and assigns a label activity in correspondence to that probability. The highest accuracy is highlighted with a bold letter.

Table 4.6: Accuracy (%) comparison of different fusion techniques on human activity datasets

Dataset	RGB+DMI Late Fusion Scores			
	Sum Fusion	Average Fusion	Concatenation Fusion	Max Fusion
SBU	98.10	96.45	96.50	98.70
MIVIA	98.40	97.20	95.75	99.41
MSR Action Pair	96.90	94.60	94.80	98.30
MSR Daily Activity	95.36	90.40	91.60	94.37

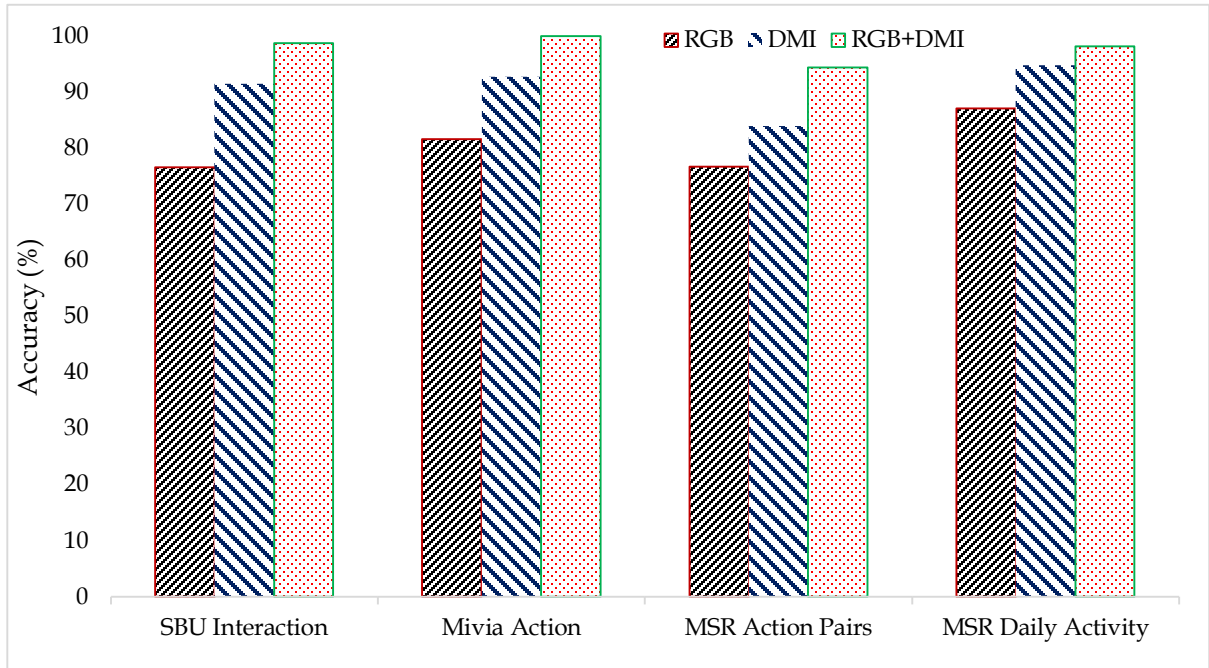


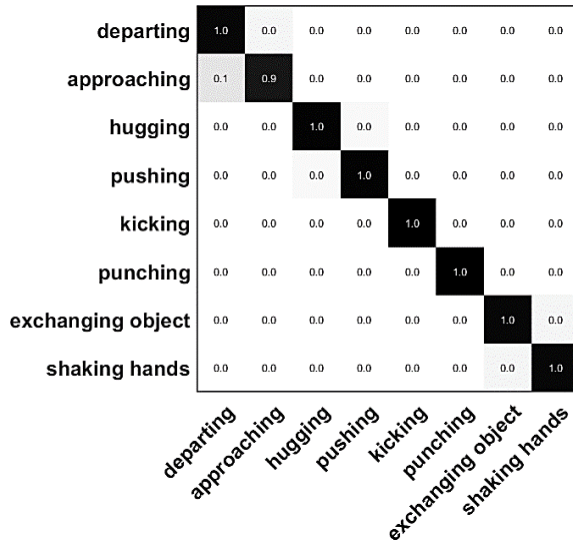
Figure 4.12: Shown results on four datasets with different data inputs: Only RGB frames, Dynamic motion image (DMI) and RGB+DMI

4.2.9 Results comparison with State-of-the-art

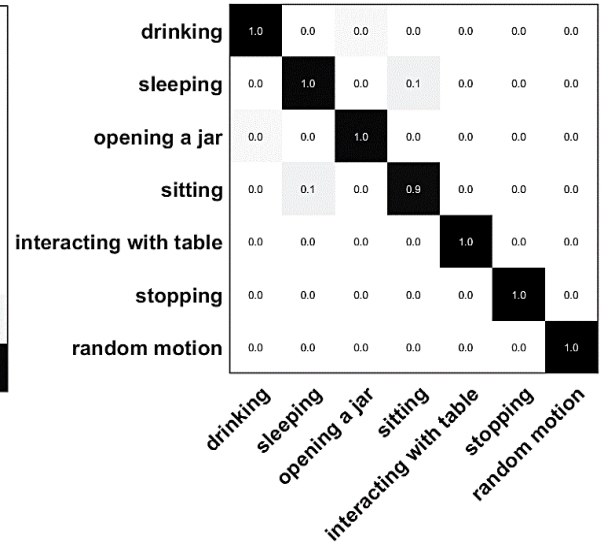
The comparison of average recognition accuracy (ARA) achieved on these datasets with two different input data stream as RGB frames, and single dynamic motion

image(DMI) and the max fusion scores of RGB with DMI is shown in Fig.4.12. It is clear from Fig. 4.12, max fusion of both RGB with DMI gives the best results as compared with independent input data streams: RGB and single DMIs. It can be seen from work Bilen et al. [183] that the dynamic image obtained with approximate rank pooling with CNN showed excellent results in many indoor or outdoor activities recognition tasks in videos. Further, features extracted from RGB only frames are not sufficient to represent the complex activities of human activities. Therefore, the proposed approach fused the discriminative features extracted from DMI and RGB frames to represent Spatio-temporal variation in activity video. The proposed model shows excellent results on all four video benchmarks. The results obtained on the MSR Daily activity dataset show somewhat less accuracy as compared with SBU Interaction, MIVIA Action and MSR Action pairs due to intra-class similarity exists between activity classes and complex background.

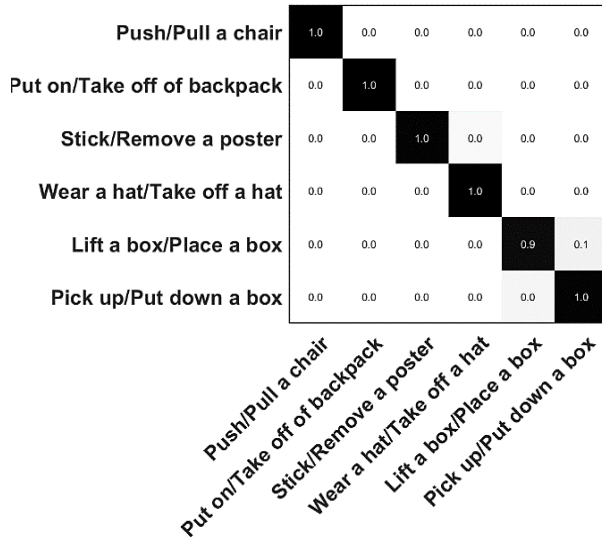
The classification result on these datasets is shown in the form of a confusion matrix as shown in Fig.4.13. The confusion matrix of SBU dataset is shown in Fig.4.13 (a). It is a challenging dataset because of the similarity of action and the same motion cues in video frames. It is clear from the confusion matrix that the main confusion exists with two similar activities such as ‘approaching’ and ‘departing’ and ‘handshaking’ and ‘exchanging object’. The state-of-the-art comparison of similar works on SBU dataset is listed in Table 4.7. Our deep model is evaluated on SBU dataset with 5-fold cross-validation similar in the work [185]. The recognition accuracy of the proposed approach with two-stream fusion shows the best accuracy on this dataset.



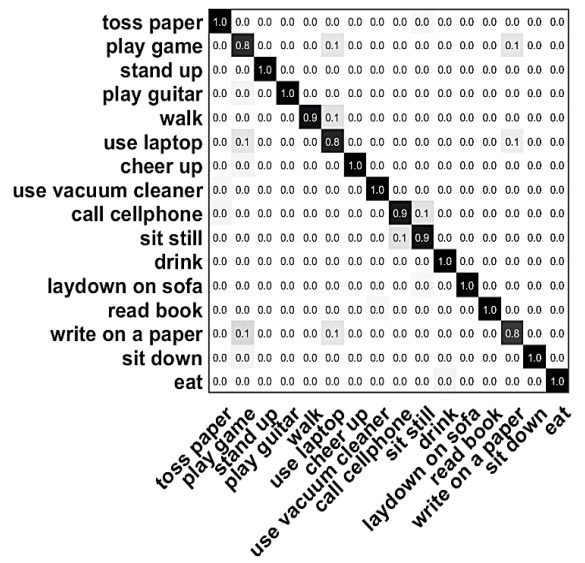
(a)



(b)



(c)



(d)

Figure 4.13: The confusion matrix of datasets: a) SBU Interaction b) MIVIA Action c) MSR Action Pairs d) MSR Daily Activity

Table 4.7: State-of-the-art comparison of SBU interaction Dataset

Works	Methods	Accuracy (%)
Yun et al. [185]	SVM and MIL Boost	87.30
Feng et al. [49]	BiVector and LSTM	82.70
Ijjina and Chalavadi [119]	ConvNet + ELM Classifier	86.58
Keçeli et al. [118]	2D and 3D ConvNet	94.70
Elboushaki et al. [122]	MultiD-CNN	97.51
Proposed Approach(RGB)	Inception-v3 + Bi-LSTM	76.60
Proposed Approach (DMI)		91.40
Proposed Approach (RGB+DMI) Max Fusion		98.70

The experimental result of MIVIA action and comparison with similar approaches are shown in Table 4.8. The LOSO cross-validation evaluation protocol is followed for training and testing the proposed model similar in [83]. It is observed from the confusion matrix shown in Fig. 4.13 (b) that actions having no motion in the video found confusion for recognize for example ‘sitting’ and ‘sleeping’ activities. The spatial features for both the activities are the same but lack of temporal cue due to no movement of actors. In this situation, DMI stream features extraction is not so useful for activity prediction. However, the proposed ConvNet model obtained the best results with the fusion of CNN-LSTM and dynamic images as compared with existing methods.

The result comparison of MSR Action Pairs is shown in Table 4.9. The evaluation protocol is followed as in similar works by Wang et al. [186]. Out of total performer, half of the actors are used for training the model and half of the actors for testing the action classes. It is observed that from the confusion matrix in Fig.4.13 (c) that all the six action pairs are correctly recognized by the proposed approach and no occurrence of intra-class pair confusion. There is slight confusion between ‘lift a box’ and ‘put down a box’ action pairs but our hybrid model shows good performance to recognized each action pairs activity.

Table 4.8: State-of-the-art comparison of MIVIA Action Dataset

Works	Method	Accuracy (%)
Carletti et al. [83]	Reject	79.80
Foggia et al. [37]	Deep Learning	84.70
Brun et al. [72]	Edit Distance	85.20
Ijjina and Chalavadi [119]	ConNnets + ELM Classifier	93.37
Saggese et al. [103]	Skeleton feature	95.00
Brun et al. [72]	Aclets sequences	95.40
Proposed Approach (RGB)	Inception-v3 + Bi-LSTM	81.59
Proposed Approach(DMI)		92.71
Proposed Approach(RGB+DMI) Max Fusion		99.41

Table 4.9: State-of-the-art comparison of MSR Action Pairs Dataset

Works	Method	Accuracy (%)
Wang et al. [186]	LOP	82.22
Jia et al. [190]	LTTL	91.40
Oreifej and Liu [187]	HON4D	93.33
Vemulapalli and Chellapa [191]	FTP representation	94.67
Ji et al. [27]	One-shot learning	95.10
Ji et al. [26]	Skeleton embedded feature	97.70
Proposed Approach(RGB)	Inception-v3 + Bi-LSTM	87.70
Proposed Approach(DMI)		94.76
Proposed Approach(RGB+DMI) Max Fusion		98.30

Table 4.10: State-of-the-art comparison of MSR Daily Activity Dataset

Works	Method	Accuracy (%)
Amor et al. [48]	Rate-Invariant Analysis	70.00
Seidenari et al. [192]	NBNN Bag-of-Poses	70.00
Cai et al. [193]	Markov Random Field	78.20
Zhang and Parker [194]	BIPOD	79.70
Ji et al. [26]	Skeleton embedded feature	81.30
Jing et al. [120]	Joint loss function	88.00
Srihari et al. [121]	4-stream CNN	89.05
Huynh-The et al. [195]	PAM + Pose-Transition	90.63
Proposed Approach(RGB)	Inception-v3 + Bi-LSTM	76.64
Proposed Approach(DMI)		83.90
Proposed Approach(RGB+DMI) Max Fusion		94.37

The confusion matrix for MSR Daily Activity dataset is shown in Fig.4.13 (d). It is seen from the confusion matrix that the main confusion occurs in similar activities such as play game using a laptop, and write on the paper. Most of the activities are correctly classified with a high confidence level. As illustrated in Table 4.10, our proposed model performed well and achieved superior accuracy with similar state-of-the-art approaches.

In the next section, a deep bottleneck multimodal feature fusion (D-BMFF) technique is proposed that utilized three different modalities RGB, RGB-D(depth) and 3D coordinates information for activity classification.

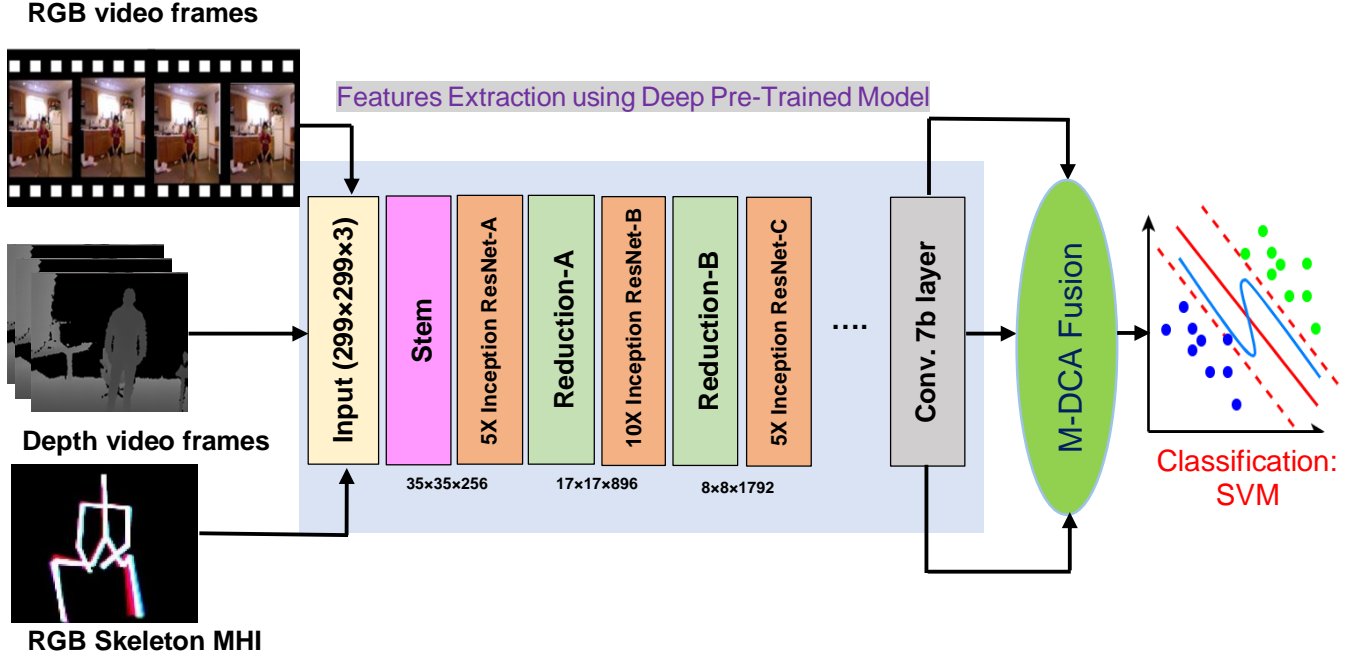


Figure 4.14: Schema of deep multimodal network based on bottleneck layer features fusion for action Recognition

4.3 A deep multimodal network based on bottleneck layer features fusion for action recognition

With the invention of advanced Kinect depth sensor various deep learning methods based on single modality (RGB, depth(D), and skeleton coordinates) and their various combinations are introduced for action recognition tasks in videos [23] [24] [25] [26] [27]. However, very few approaches are based on the combination of RGB, depth and 3D-skeleton coordinates for activity recognition [28]. Such activity recognition solutions showed promising results but had limitations due to positions of joints and coordinates of upper or lower body parts used to represent activity [48] [49]. Sometimes skeleton data is degraded due to noise and occlusions present in RGB-D images. Therefore, these single modalities based solutions are less effective for action representation. In the proposed work, we utilized three modalities RGB, depth, and 3D coordinates information for activity classification because it helps for better recognition and complete utilization of information

available from a depth sensor video simultaneously. The details descriptions of the proposed model are explained in the following sub-sections

4.3.1 Deep fusion framework for Human Activity Recognition

This is known fact that a new deep model requires a large amount of training data and computational power when trained from raw data. Consequently, it is not a feasible task for collecting and labelling a large amount of data which make a deep learning model quite difficult to apply. Due to constrained of large training data, we utilized transfer learning for human activity recognition. In the proposed deep framework, we have fine-tuned the bottleneck layers and fully connected (FC) layer of the pre-trained inception-Resnet-v2 architecture [196] for different input streams. The RGB and depth frames activity videos are fed to the pre-trained network at regular interval of ten frames for spatial features extraction while a single Skel-MHI image for each activity is given as an input for additional temporal features evidence. Further, the extracted multimodal features from three data streams are fused at (Conv '7b') and (FC) layers using Multiset DCA fusion technique. The activity classification is carried through the SVM machine learning technique after fusing the features obtained from a pre-trained deep network. Fig. 4.14 depicted the underlying architecture of the proposed framework for human activity recognition.

4.3.1.1 Features Extraction

It is experimentally observed that deep pre-trained model trained on a large annotated dataset is exchangeable to action recognition task with the smaller training dataset. It can be observed from table 4.11 that deep Inception architecture is shown promising results at little computational cost as compared with other deep architecture. This is because by adding a residual unit with the traditional inception model given a state-of-the-art performance in recent ILSVRC challenges [196].

Further, it has increased the training speed of Inception model by a sufficient margin and improve the image recognition task significantly.

Table 4.11: Comparison of accuracies of Inception-Resnet-v2 architecture with similar state-of-the-art architectures

Deep Pre-trained Network	Crops	Top-1 Error	Top-5 Error	Filter bank sizes
Inception-v3 [180]	144	18.9%	4.3%	-
Inception-ResNet-v1 [196]	144	18.8%	4.3%	$k=192, l=224, m=256, n=384$
Inception-v4 [196]	144	17.7%	3.8%	$k=192, l=192, m=256, n=384$
Inception-Resnet-v2 [196]	144	17.8%	3.7%	$k=256, l=256, m=384, n=384$

In this work, we used Inception-ResNet-v2 architecture for features extraction from input videos. This deep model is trained on a large image database which showed excellent results with minimum error on image classification competitions. It is a 164 layers deep network that accepts $299 \times 299 \times 3$ size of input images. It contains two deep architectures residual connections [197] and the latest inception architecture Inception-v4 [196]. It is observed that residual connections are a favourable choice for training very deep architectures and so that computational efficiency of deep Inception would be increased by adding the residual units into the architecture. The details structure of each Inception-ResNet blocks A, B and C is shown in Fig.4.15. The three blocks of inception block is followed by a filter expansion layer (1×1 convolutional layer without activation). This 1×1 Conv layer help for scaling the dimensionality of the filter bank earlier the addition to match the depth of the input.

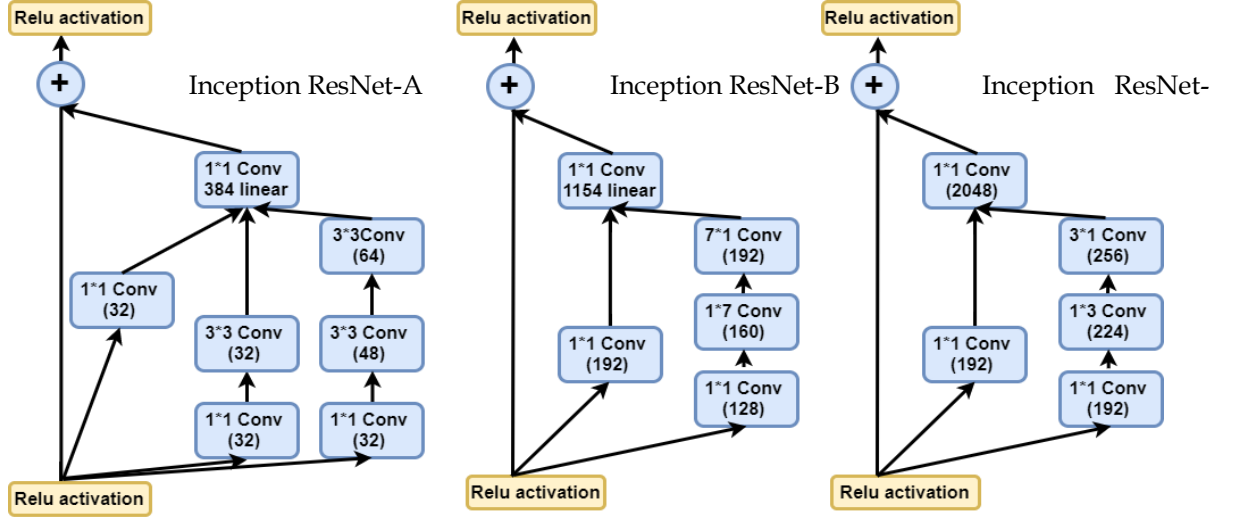


Figure 4.15: Shown the systematic block diagram for 35x35 grid module of Inception ResNet-A, 17x17 grid module of Inception ResNet-B, and 8x8 grid module of Inception ResNet-C

4.3.1.2 RGB Skeleton Motion History Images (RGB-SklMHI)

It is experimentally observed that the skeleton joint coordinates are not capable enough to discriminating some activities due to noise and occlusion errors such as self-occlusion with body parts etc. Sometimes skeleton data is degraded due to noise and occlusions present in RGB-D images and less effective for action recognition. Further, it is observed that binary MHI images are not sufficient enough to discriminate some similar motion pattern such as 'sitting down' and 'standing up' position. To overcome such limitations the concept of a RGB skeleton MHIs from the 3D coordinates is introduced in the work of [198]. The RGB skeleton MHIs are obtained from joints coordinates based on averaging the binary MHIs on the sequential time interval for each action classes. These colour MHIs are based on view temporal template and robust to shadows, noise, occlusion and illuminance conditions for action recognition tasks. The formation of RGB-SklMHI from 3D coordinates joint information is based on the sequential time interval for each action classes are described in Algorithm 1.

Algorithm 1: Colorization of Skeleton Motion History Image
Input: 3D joints coordinates
Output: RGB Skeleton MHIs
<p>Input:</p> <ol style="list-style-type: none"> 1. $N=1,2,3,\dots,n$; total joints 2. $f=1,2,3,\dots,n$; total frames 3. $p=1,2,3,\dots,n$;all labels 4. $I(X, Y)$ template RGB-SklMHI 5. $I(\text{SkeletonConnectionMap}) = [\text{spine}, \text{left}, \text{right}, \dots]$ 6. $I(\text{ColorSet}) = [0,1,0;0,1,0.33, \dots]$ 7. $\text{RGBSkeletonImg}()$ <p>Procedure: $\text{Create_SklImg}((N, p, f, I(X, Y), \text{ColorSet}, \text{color_c}, \text{SkeletonConnectionMap}))$</p> <p>for $p \leftarrow \text{all labels}$ (do)</p> <p>for $f \leftarrow \text{all frames}$ (do)</p> <p>for $\text{frame} = \text{startframe} + (N-1) : \text{endframe}$</p> <p>$I(X, Y) = \text{zeros}(\text{height}, \text{width}, 3);$</p> <p>$\text{color_c} = 1;$</p> <p>for $k = \text{frame} - (N-1) : \text{frame}$</p> <p>$\text{color_img} = (\text{ColorSet}, \text{color_c}, \text{SkeletonConnectionMap});$</p> <p>$I(X, Y) = I(X, Y) + \text{color_img};$</p> <p>$\text{color_c} = \text{color_c} + 1;$</p> <p>end for</p> <p>end for</p> <p>end for</p> <p>end procedure</p> <p>Output: RGB Skeleton MHIs.</p>

We have generated the RGB-SklMHI for each RGB-D datasets using the procedure explained Algorithm 1. The RGB-SklMHI are given as an input to the pre-trained network for extracting the temporal variations patterns in an activity video. The classification results obtained using individual stream and fusion with other streams are discussed in section 4.3.5.

4.3.1.3 Multimodal features fusion using Discriminate Correlation Analysis (M-DCA)

Deep trained features vector contained the highest redundant information with high dimension. Therefore, it is required to adopt an effective dimension reduction

technique as compared to conventional data reduction technique. Discriminant correlation analysis (DCA) fusion technique is proposed by Haghighat et al. [199] for real-time multimodal biometric recognition system with low computational complexity. It is a compelling fusion technique which maximized the pairwise correlations among the features sets. Further, it removes the between class correlation and restricting the correlation to be within classes. It can be useful for fusing the features extracted the multiple modalities or combining the different feature vectors extracted from single modalities.

DCA algorithm fused the multiple feature sets from multi-modal data streams at the decision level. It showed promising results with more discriminative efficiency. In proposed work, we have fused the features obtained from three different inputs by pre-trained deep architecture at the bottleneck layer ‘Conv 7b’ before the fully connected layers. Further, the representation ability further increases as compared fusion at fully connected layer. Algorithm 2 shows the fundamental steps for fusing the feature sets using DCA technique.

In this work, we used the multiple features sets DCA for fusing the features extracted from three input streams RGB, RGB-D and RGB-SkelMHI. Let, we have m set of features, $\mathcal{X}_i \in \mathbb{R}^{i \times n}, i = 1, 2, 3 \dots \dots m$, which are selected by their rank, i.e. $rank(\mathcal{X}_1) \geq rank(\mathcal{X}_2) \dots \dots \geq rank(\mathcal{X}_m)$. Since we have three sets of features vector obtained from the pre-trained model. Therefore, DCA is applied for two sets of feature stream at a single time, i.e. features vector from the RGB, RGB-D and Skel-MHI input data streams are fused first or which are having highest rank and next to the highest rank feature set will be fused. The length of fused feature vector r are selected as given as:

$$r \leq \min(c - 1, rank(\mathcal{X}_i), rank(\mathcal{X}_j)) \quad (4.20)$$

Fig.4.16 depicted an example of multiset DCA fusion for three data streams RGB, RGB-D and RGB-SkelMHI feature sets.

Algorithm 2	Multi-set features fusion using DCA
Step 1	<p>The mean \tilde{m}_i and \tilde{m} is calculated for each class and training data, respectively as:</p> $\tilde{m}_i = \frac{1}{n_i} \sum_{j=1}^{n_j} m_j^i, \quad \tilde{m} = \frac{1}{n} \sum_{i=1}^c n_i \tilde{m}_i$ <p>Where, n number of images, c number of classes.</p>
Step 2	<p>The covariance matrix is computed as $C = \Phi^T \Phi$</p> <p>where $\Phi = (\sqrt{n_1}(\tilde{m}_1 - \tilde{m}), \dots, \sqrt{n_c}(\tilde{m}_c - \tilde{m}))$,</p>
Step 3	Calculate the SVD of C as $C = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T$, $\mathbf{\Lambda} = \text{diag}(\lambda_1 \lambda_2, \dots, \lambda_c) (\lambda_1 \geq \lambda_2, \dots \geq \lambda_c)$
Step 4	The transform matrix is calculated as $\mathbf{P} = \Phi \mathbf{U}_r \mathbf{\Lambda}_r^{-\frac{1}{2}}$.
Step 5	Repeat the steps (1-4) to compute the transform matrix for \mathcal{X}_{RGB} , \mathcal{X}_{RGB-D} , and $\mathcal{X}_{Skel-MHI}$ streams separately.
Step 6	Calculate the transform data: $\mathcal{Z}_1 = \mathbf{P}_1^T \mathcal{X}_{RGB}$, $\mathcal{Z}_2 = \mathbf{P}_1^T \mathcal{X}_{RGB-D}$, and $\mathcal{Z}_3 = \mathbf{P}_1^T \mathcal{X}_{Skel-MHI}$.
Step 7	Evaluate the between-set covariance matrix for two sets of the transformed feature set: $\mathcal{S}_{bwn} = \mathcal{Z}_1 \mathcal{Z}_2^T$.
Step 8	Evaluate the SVD of $\mathcal{S}_{bwn} = \mathbf{V} \mathbf{\Sigma} \mathbf{V}^T$
Step 9	Define the transform matrix: $\mathcal{T} = \mathbf{V} \mathbf{\Sigma}^{-1/2}$
Step 10	Evaluate the transformed data: $\mathcal{X}'_1 = \mathcal{T}^T \mathcal{Z}_1$, $\mathcal{X}'_2 = \mathcal{T}^T \mathcal{Z}_2$, and $\mathcal{X}'_3 = \mathcal{T}^T \mathcal{Z}_3$,
Step 11	Apply the Multiset DCA on two feature sets at a time using the maximum length of the fused feature vector r according to Eq. (4.20). as shown in Fig. (4.13).

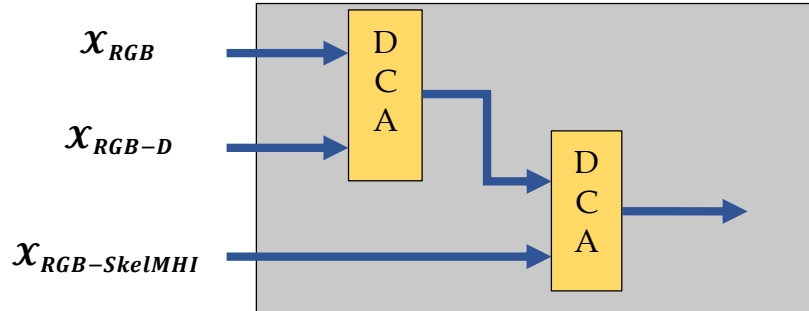


Figure 4.16: Shown the multiset DCA analysis on three data stream features RGB, depth and RGB-SkelMHI

4.3.2 Implementation Details

We have trained and test our model on MatConvNet toolbox on a system having an NVIDIA TITAN RTX memory card with Xenon processor. The RGB and depth videos with a batch size of 8 videos are fine-tuned on Conv ‘7b’ up to 140 epochs with a dropout of 0.8 after average pooling layer to avoid the overfitting. Initially, we used the momentum constant of 0.9 and a learning rate of 0.0001. The scaling factor is choosing between 0.1 to 0.3 for residual connection to stabilize the training data. For the transparency of results and comparison, we adopt the same evaluation protocol for each dataset as in their original work. The Leave one out cross-validation(LOOCV) training protocol used for UT Kinect [200], CAD 60 [201], and Florence 3D Action [192] datasets. The 5-fold cross-validation evaluation criteria applied for SBU Interaction dataset. The performance of the proposed method is measure in terms of Average Recognition Accuracy (ARA) per class for many classes (\mathcal{C}_i), which is calculated as in Eq.(4.17).

4.3.2.1 Human Activities RGB-D Datasets

In order to test the performance of the proposed method, evaluation is done over four publically available RGB-D datasets: UT Kinect, CAD 60, Florence 3D, and SBU Interaction. Our deep multimodal fusion approach shows superior recognition accuracy as compared with earlier approaches on such dataset. The detailed information about the RGB-D datasets is explained in the following subsection.

4.3.2.2 UT Kinect Action Dataset

This dataset is introduced by Xia et al. [200] and captured through a single Kinect sensor with three different modalities color, depth and 3D skeleton coordinates simultaneously. It consists of 10 actors performing 10 action class: waving, walking, sitting down, sitting up, carry, picking up, pulling, clapping hands, throwing and pushing. There are 200 video sequences with total 6220 frames. The spatial

resolution for RGB and depth maps are 640×480 and 320×240 pixels respectively and recorded with 30 frames per second. The skeleton information is capture with 20 skeleton joints. It is a challenging dataset in terms of intra-similar activities. The sample images of UT dataset are shown in Fig.4.17.

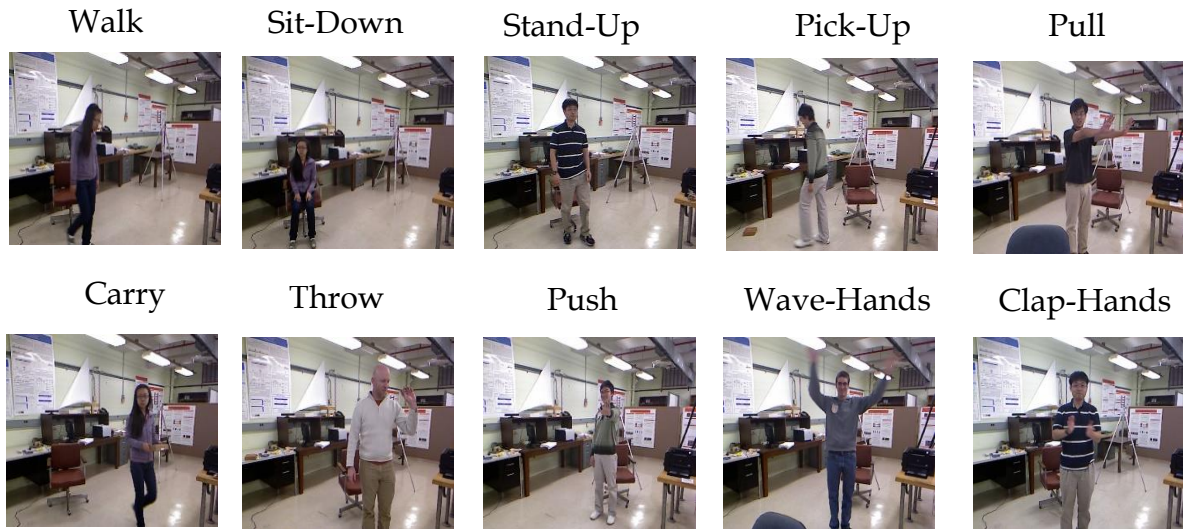


Figure 4.17: Sample frames from UT Kinect dataset

4.3.2.3 CAD 60 Action Dataset

Sung et al. [201] proposed this dataset to recognized the human activity in the various indoor home environment living room, bathroom, office, kitchen and bedroom. This dataset is recorded by the Kinect sensor with three different modalities such as RGB, Depth and 3D skeleton coordinates data. There are four actors 2 males and 2 females performing 12 different activities: working on computer, writing on whiteboard, relaxing on couch, talking on couch, cooking (chopping), opening pill container, cooking (stirring), talking on phone, wearing contact lens, drinking water, brushing teeth, rinsing mouth. The RGB-D frames have the spatial resolution 640×480 and skeleton joints contain 15 body joints information data. The sample images from this dataset are depicted in Fig.4.18.

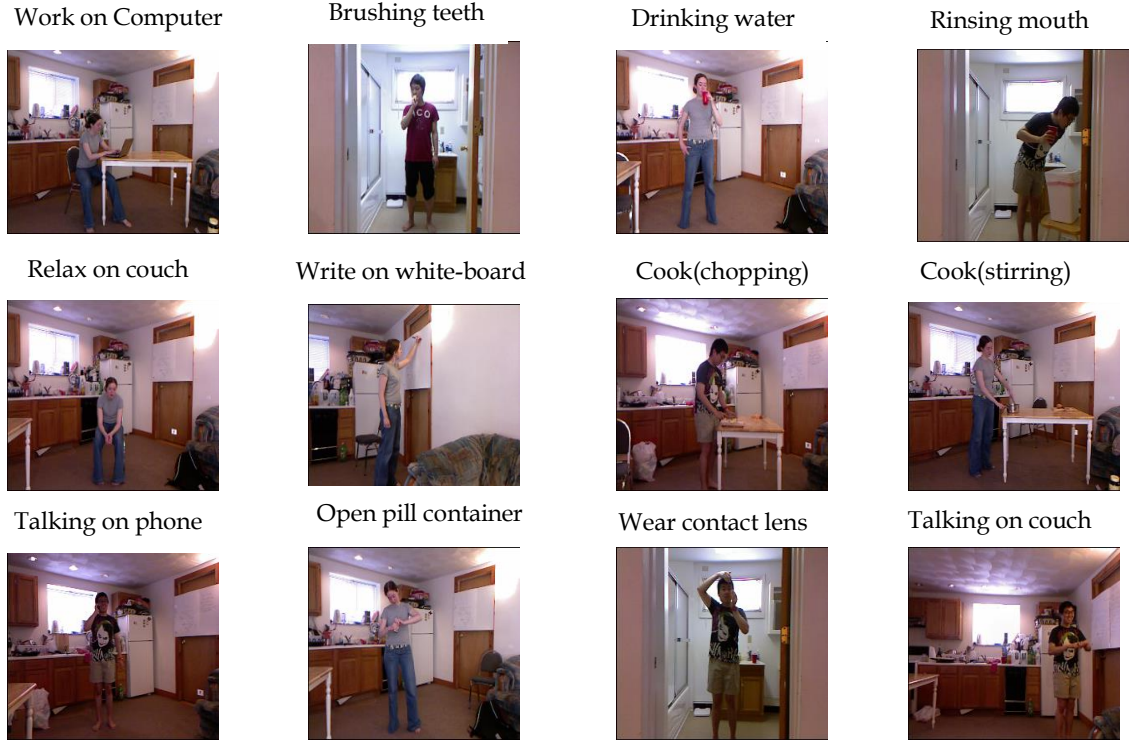


Figure 4.18: Sample frames from CAD 60 Dataset

4.3.2.4 Florence 3D Action Dataset

Seidenari et al. [192] proposed this dataset at the University of Florence. It has been captured through Kinect Sensor. There are 10 actors performing 9 human activities: stand up, sit down, bow, tight lace, wave, answer phone, read watch, clap, drink from a bottle. Each actor repeated above activities two to three times. There is a total of 215 activity video sequences. The sample images from this dataset for various activities is depicted in Fig.4.19.

4.3.3 Result Analysis

The confusion matrices for all four RGB-D datasets are depicted in Fig. 4.20. We can see that the proposed method performs well on all the datasets. The multi-modal bottleneck layer fusion in conjunction with SVM classifier significantly increases the classification accuracy. The confusion matrix for UT Kinect dataset is shown in Fig. 4.20 (a).

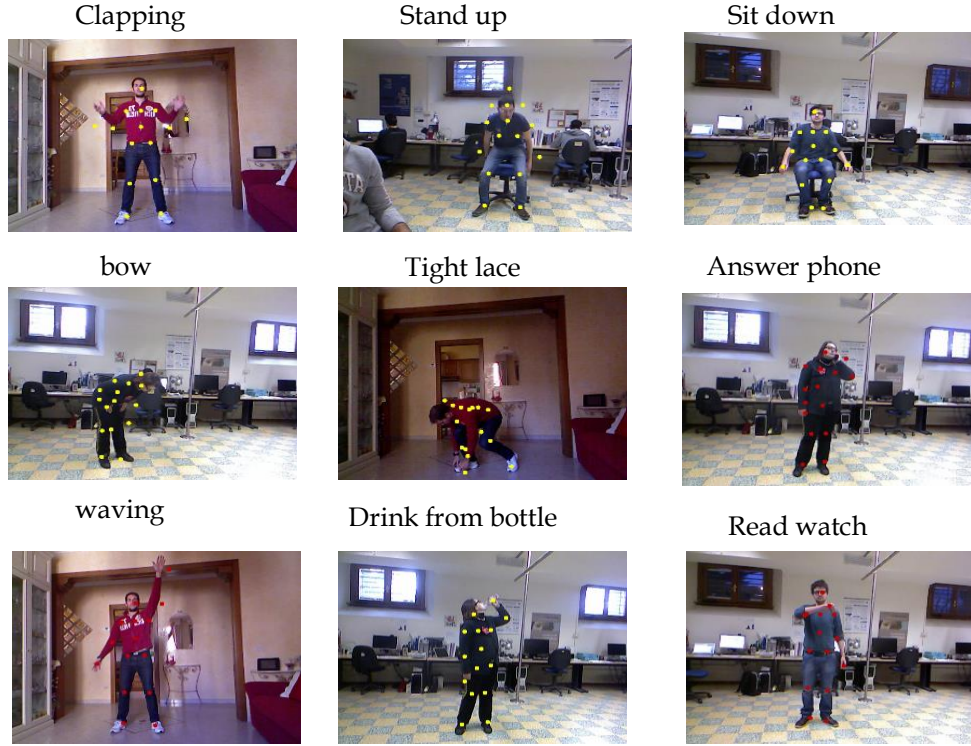
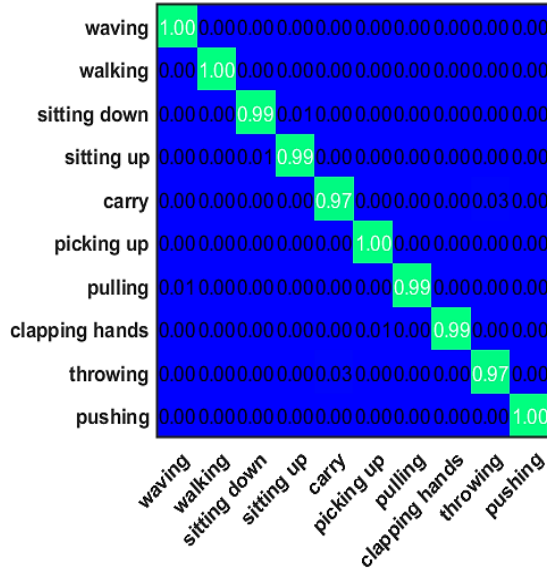


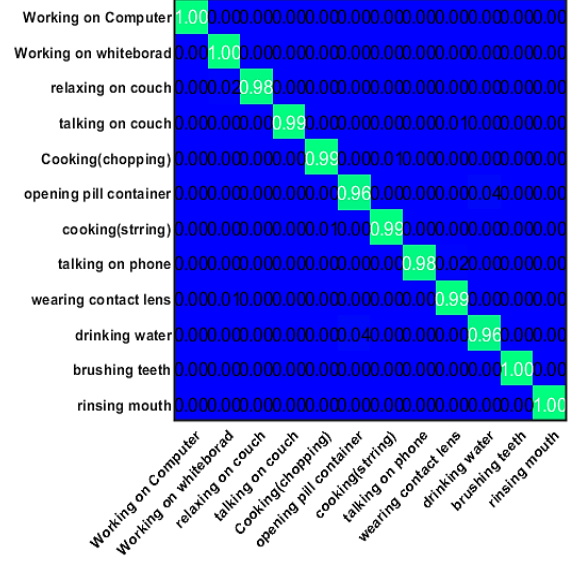
Figure 4.19: Sample frames from Florence 3D Action

It is noted that most of the activities in the dataset are classified with higher confidence by the proposed D-BMFF technique. Still, there is slight confusion between the ‘throwing’ and ‘carry’ activities due to the occlusion caused by the human-object interaction and field of view in which captured by the sensor. However, the model demonstrates the highest accuracy as compared with other existing solutions.

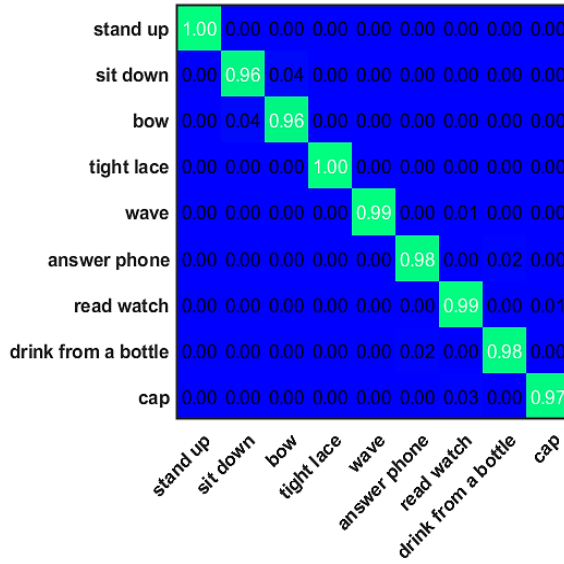
The confusion matrix for CAD-60 dataset is shown in Fig.4.20 (b). It is challenging dataset because most of the activities consist of the upper body part movement only. Most of the activities are classified, but there is slight confusion for actions such as: ‘opening pill container’, ‘drinking water’, ‘relaxing on the couch’, and ‘talking on the phone’. This dataset action video shows high semantic level activities that could be considered as an advantage such as drinking, eating.



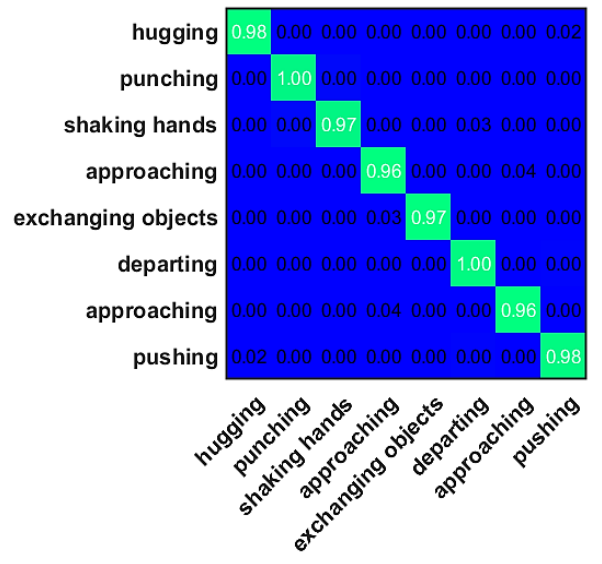
(a)



(b)



(c)



(d)

Figure 4.20: Shown the confusion matrix of proposed D-BMFF technique on four datasets: a) UT Kinect, b) CAD-60, c) Florence Action 3D, and d) SBU Interaction dataset

The confusion matrix for Florence 3D dataset is shown in Fig.4.20 (c). We can see that the proposed method performs well and shows better results than existing works on all actions of this dataset. The main confusion occurs for actions like ‘sit down’, ‘drinking water’ and ‘answering the phone’. It is because of the same action is performing in various ways and human-object interaction.

The confusion matrix for SBU Interaction dataset is shown in Fig.4.20 (d). It is challenging dataset because the interaction activities performed by the pair of actors, unlike the single actor independently. The actions are non-periodic and very similar in movements. We can see that the main misclassification is observed in ‘shaking hands’ and ‘exchanging the object’ or ‘approaching and departing’ activities of this dataset. Besides this high similarity actions, the proposed method shows superior performance on all activities.

4.3.4 State-of-the-art Comparison

In this section, we discussed and compared the results obtained by the deep model on four RGB-D datasets. The average recognition accuracies on UT Kinect, CAD-60, Florence 3D, SBU Interaction are: 99%, 98.50%, 98.10%, and 97.75% respectively. Table 4.12 listed the results obtained by different existing works on UT Kinect dataset. We can see that the proposed method shows the comparable accuracy on this dataset as compared with state-of-the-art approaches. It can be observed that most of the works utilize only skeleton input for action representation and very few on multiple modalities [25] [202]. However, it seems easy to implement the model using skeleton information only, but they suffer due to self-occlusion and noise errors present to estimates the skeleton coordinates. Our method shows higher activities than the cited works [203] [106] [204] by a sufficient margin. The accuracy is comparable with the work [25] as they fused the depth and skeleton information for action recognition.

Table 4.12: Comparison of multimodal features fusion score using Inception-Resnet-v2 on UT Kinect Dataset

Works	Input	Method	Accuracy (%)
Avola et al. [202]	RGB + Depth	BoW + ELM	84.00
Devanne et al. [203]	Skeleton	Riemannian Manifold	91.50
Ghodsi et al. [106]	Skeleton	Joint skeleton trajectories	96.80
Zhang et al. [204]	Skeleton	Multilayer LSTM	97.00
Vemulapalli et al. [205]	Skeleton	Lie Group	97.08
Huynh-The et al. [195]	Skeleton	PAM + Pose-Transition	97.00
Nguyen et al. [25]	Depth +skeleton	KM-IELLogE-IELLogE	99.50
Proposed	RGB+ Depth+ Skeleton	D-BMFF	99.00

Table 4.13 depicted the state-of-the-art comparison of CAD-60 dataset. It is noted that the proposed model gave the highest accuracy as compared with similar approaches. Our model shows a significant hike in accuracy as compared with the multimodal features based works [28] [109] [110].

Table 4.13: Comparison of multimodal features fusion score using Inception-Resnet-v2 on CAD 60 Dataset

Works	Input	Method	Accuracy(%)
Avola et al. [202]	RGB+ Depth	BoW +ELM	82.60
Sung et al. [201]	Depth +skeleton	Max Entropy Model	83.20
Raman and Maybank [110]	Depth+ Skeleton	H-HMM	85.40
Khaire et al. [28]	RGB+ Depth+ Skeleton	5-CNNs	90.00
Kong et al. [109]	RGB+ Depth+ Skeleton	CMFL	94.10
Liu et al. [206]	Skeleton	Pose+ kNN	95.75
Li et al. [207]	Skeleton	ShapeDTW	97.30

Franco et al. [144]	RGB+ Skeleton	HOG+BOW fusion	98.30
Proposed	RGB+ Depth + Skeleton	D-BMFF	98.50

Table 4.14: Comparison of multimodal features fusion score using Inception-Resnet-v2 on Florence 3D Action Dataset

Works	Input	Method	Accuracy (%)
Seidenari et al. [192]	Skeleton	NBNN Bag-of-Poses	82.00
Devanne et al. [203]	Skeleton	Riemannian Manifold	87.04
Vemulapalli et al. [205]	Skeleton	Lie Group	90.88
Salih and Youssef [208]	Skeleton	STIP MSH	86.13
Huynh-The et al. [195]	Skeleton	PAM + Pose-Transition	92.09
Sun et al. [209]	Skeleton	Local and global Histogram	92.19
Yang et al. [107]	Skeleton	Latent Max-Margin	93.42
Nguyen et al. [25]	Skeleton +Depth	KM-IELLogE-IELLogE	95.37
Proposed	RGB+ Depth+ Skeleton	D-BMFF	98.10

The results obtained on Florence 3D dataset are shown in Table 4.14. It is observed that our multi-modal approach shows higher accuracy despite the intra-class similarity and different way of performing the action classes. The state-of-the-art comparison of SBU interaction dataset is shown in Table 4.15. The proposed model outperforms the other existing approaches to this challenging dataset. We have achieved the highest accuracy of 97.75% on such complex interaction activity dataset. Our model gained a sufficient margin accuracy as compared with only skeleton features based activity descriptors [49] [185].

4.3.4.1 Accuracy Comparison with Existing Pre-Trained Deep Models

In this section, we discussed the recognition accuracies using both single and multi-modal fusion streams obtained through different pre-trained deep architectures. The multimodal features are extracted by fine-tuned the pre-trained inception-ResNet-v2 model and fused the extracted features from three different data streams at ‘Conv 7b’ layer just before the fully connected layers. It is empirically observed that fine-tuning on ‘Conv 7b’ layer showed better results than features extraction from a fully connected layer in conjunction with multi-class SVM classifier. This is because SVM classifier shows better results for intermediate layers for small datasets as compared with features extracted from fully connected layers.

Table 4.15: Comparison of multimodal features fusion score using Inception-Resnet-v2 on SBU Interaction Dataset

Works	Input	Methods	Accuracy (%)
Yun et al. [185]	Skeleton	SVM and MIL Boost	87.30
Feng et al. [49]	Skeleton	Bi-Vector and LSTM	82.70
Ijjina and Chalavadi [119]	RGB + Depth	ConvNet + ELM	86.58
Keçeli et al. [118]	Depth	2D and 3D ConvNet	94.70
Khaire et al. [28]	RGB + Depth + Skeleton	VGG-F+WPM	96.26
Proposed	RGB+ Depth+ Skeleton	D-BMFF	97.75

Furthermore, the features are extracted by fine-tuning at fully connected layers. The obtained results compare with the bottleneck layers’ results. Table 4.16 depicts the result of the different pre-trained model on UT Kinect dataset. The fusion results obtained at Conv ‘7b’ layer is 6% higher than ‘FC’ layer by using Inception-ResNet-v2 architecture. For a comparison point of view, we have evaluated the proposed model on existing famous pre-trained models. We have fine-tuned the action classes on other deep architectures at the last fully connected layers for Inception- v1 [180], ResNet [197], and VGGNet [210].

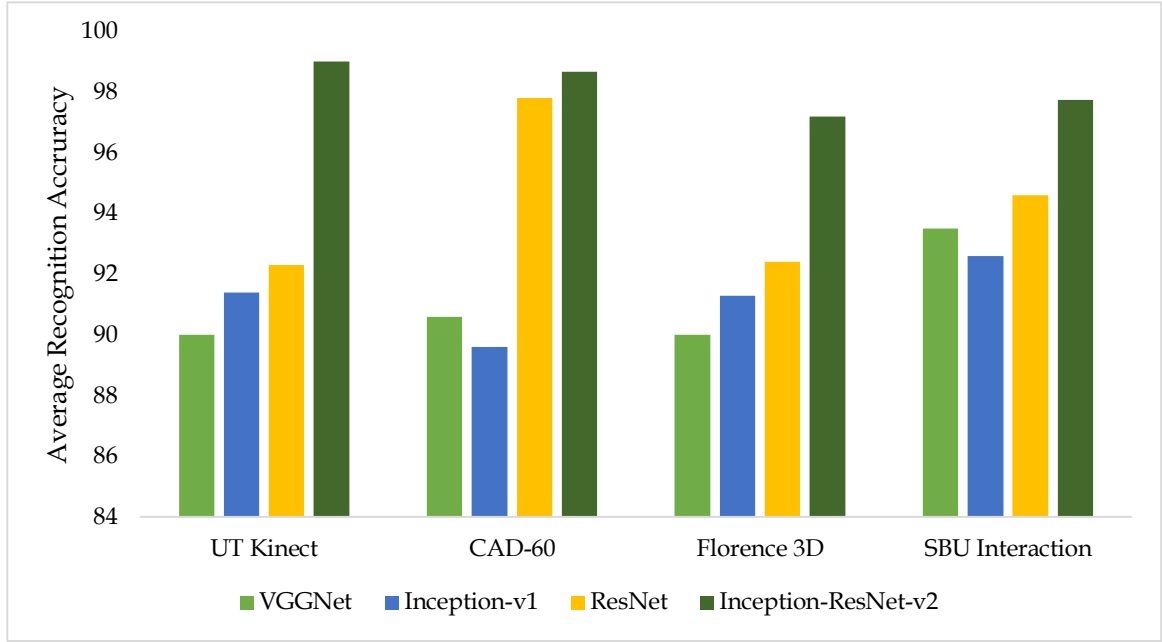


Figure 4.21: Shown the fusion accuracies comparison obtained from different pre-trained models on RGB-D datasets

Similarly, Tables 4.17, 4.18, and 4.19 listed the results obtained from fine-tuned deep architectures for CAD-60, Florence Action, and SBU Interaction datasets, respectively. The fusion accuracy increase for CAD-60(8%), Florence 3D (6%), and SBU (7%) datasets. It is observed that fusion results are increased by a significant margin when the features are extracted from intermediate Conv '7b' layer as compared with the fully connected layer.

Table 4.16: Results comparison of Pre-trained architectures on UT Kinect Dataset

Model Dataset	Modality	VGGNet [210] (FC)	Inception- v1 [180] (FC)	ResNet [197] (FC)	Inception-ResNet-v2	
					(FC)	(Conv '7b')
UT Kinect	RGB	84.50	80.70	85.40	86.00	90.70
	Depth	82.50	85.10	86.20	87.70	91.20
	Skeleton	86.40	88.50	89.40	86.60	90.10
	Fusion	90.00	91.40	92.30	93.60	99.00

Table 4.17: Results comparison of Pre-trained Architectures on CAD-60 Dataset

Model Dataset	Modality	VGG19 [210] (FC)	Inception- v1 [180] (FC)	ResNet [197] (FC)	Inception ResNet-v2	
					(FC)	(Conv '7b')
CAD 60	RGB	85.80	87.00	86.10	89.30	94.40
	Depth	84.40	88.70	89.60	85.70	91.30
	Skeleton	89.10	88.50	91.10	89.80	95.60
	Fusion	90.60	89.60	97.80	90.10	98.66

Table 4.18: Results comparison of Pre-trained Architectures on Florence 3D Dataset

Model Dataset	Modal- ity	VGG19 [210] (FC)	Inception-v1 [180] (FC)	ResNet [197] (FC)	Inception ResNet-v2	
					(FC)	(Conv '7b')
Florence 3D	RGB	86.50	85.60	87.70	90.60	92.90
	Depth	84.60	88.10	89.60	87.40	91.20
	Skeleton	86.70	92.30	90.50	88.20	93.40
	Fusion	90.00	91.30	92.40	91.35	97.20

Table 4.19: Results comparison of Pre-trained Architectures on SBU Interaction Dataset

Model Dataset	Modal- ity	VGG19 [210] (FC)	Inception- v1 [180] (FC)	ResNet [197] (FC)	Inception ResNet-v2	
					(FC)	(Conv '7b')
SBU	RGB	90.80	90.00	92.80	92.40	96.80
	Depth	89.50	89.80	90.10	89.30	92.30
	Skeleton	91.40	90.50	92.10	87.70	95.60
	Fusion	93.50	92.60	94.60	90.70	97.75

It is observed from Tables (4.16-4.19) that skeleton streams given the maximum accuracy for each dataset as compared with RGB and depth frames. This is because 3D coordinates are first converted into motion history skeleton images for each activity videos. The RGB-SkelMHIs incorporates temporal pose variations and robust to variations due to similar action classes.

4.4 Significant Outcomes

The deep learning-based architectures are dominated and widely adopted for computer vision applications, especially in visual recognition. The above-mentioned study proposed the two-deep learning-based ConvNet architectures are presented to overcome the limitations of handcrafted solutions. After doing this empirical study, we observed the following significance outcomes as:

- It is experimentally observed that deep pre-trained model trained on a large annotated dataset is exchangeable to action recognition task with the smaller training dataset.
- The first deep ConvNet model utilized two parallel deep learned architectures, first is the pre-trained CNN and Bi-LSTM to extracted the spatial features from the RGB frames and on the other hand the second stream used a pre-trained CNN fine-tuned with fully connected layers to extracted the temporal features from DMIs.
- It is observed that the activities which do not have motion, the CNN-Bi-LSTM stream combination classify the activity classes with better recognition accuracy. On the other hand, the activities which have high motion, the dynamic images are used to boost the prediction with the CNN-LSTM stream after late fusion at softmax layer.
- It can be observed that from Tables 4.7,4.8,4.9 and 4.10, the comparisons with other state-of-the-art are outlined for the proposed deep architecture

proving the dominance of the framework in terms of accuracy. The prediction accuracy is computed on four publically available video benchmarks as SBU (98.70%), MIVIA (99.41%), MSR Action Pairs (98.30%), and MSR Daily Activity (94.37%).

- It is observed that single modality based approaches severe affected by self-occlusions, noise and error present in the video, especially in skeleton coordinates. Therefore, the multimodal feature fusion increases the recognition rate and helps to complete utilization of available data.
- In the second approach, we proposed a deep bottleneck multimodal features extraction (D-BMFF) technique for human activity recognition by utilized the all three modalities RGB, Depth(d), and 3D Skeleton information.
- It is noted that from Tables, 4.12,4.13,4.14, and 4.15, multiset features fusion at the bottleneck layer before the top layer increases the classification accuracy with the SVM classifier technique.
- The proposed model is evaluated on four RGB-D datasets and outperforms the state-of-the-art approaches. It achieved the highest accuracy on CAD-60, Florence 3D and SBU interaction dataset and comparable performance on UT Kinect dataset.
- However, multimodal features extraction from intermediate layers is more complex due to a large number of deep layers of the pre-trained architectures. In future, we will try to developed deep model by utilizing the parallel computation and representation of CNN architecture.

Chapter 5

Conclusion and Future Scope

This chapter provides a summary of proposed works, significant finding, contributions and limitations. Further, we also suggest some future directions, short-term and long-term perspectives for human activity recognition in videos.

5.1 Conclusions

This work started with background knowledge about human action or activity recognition in the video sequences, and major challenges existing for action recognition in videos, associated HAR applications and their potential manifestation by using existing solutions. Based on the theoretical and experimental works in this study, we can draw the following conclusions are as follows:

- It has been observed that to recognized action in single still image is more challenging than video sequences. It can be considered more challenging to recognised action in still image than video analysis. Because it does not involve the temporal variations, illumination variation and alignment of the images.
- In this context, a multiresolution based feature descriptor model (EMRFs) is developed for the recognition of human action in video sequences with the help of still key pose images.
- It is observed that it not a good idea to extract key poses frames using normal distance function or some fixed threshold because of the risk of high information lost. Therefore, still key poses images are selected from videos

using fuzzy inference model based on maximum histogram distances between adjacent frames that removed adjacent frames redundancy.

- Furthermore, the Gabor wavelet transform is used to make these key poses frames invariant for different orientations and scales. The parameters such as numbers of bins, number of scale and orientations are chosen empirically for development of the EMRFs.
- The performance of the EMRFs is measured on publically available such as Weizmann, KTH, UCF YouTube datasets, those are challenging in respect of lightening variations, and zoom in and out.
- The developed handcrafted feature descriptor showed best results as compared with state-of-the-art approaches.

In the second part of this thesis work, we proposed two deep learning-based architectures for activity recognition and concluded that:

- It is observed that the improved image recognition approaches extend human action recognition in video sequences. Human action motion in a video can be disintegrated into spatial and temporal features. The spatial features contain the appearance information about an object in each video sequence while temporal features represented in the form object moving across the video sequences.
- The first-deep model architecture, we utilized two parallel stream deep architectures, first upper stream combination of the pre-trained CNN and Bi-LSTM to extract the spatial features from the given RGB frames. The second lower stream is a pre-trained CNN fine-tuned with fully connected layers fed with DMIs as input for temporal extraction cues. Therefore, a robust two-stream deep ConvNet model is developed for the recognition of single, multi-person and human-object interaction (HOI) activities in the video sequence.

- The proposed ConvNet is evaluated on four publically available standard video benchmarks: SBU Interaction, MIVIA Action, MSR Action Pairs, and MSR Daily Activity.
- These datasets are challenging due to the existence of non-periodic action, human-object interactions, intra-class similarity, and similar motion cues activities. The comparisons of accuracy with state-of-the-art and outperform, proving the dominance of the proposed framework.
- In the second-deep model architecture, a deep bottleneck multimodal features extraction (D-BMFF) technique for human activity recognition that utilized three modalities RGB, Depth(D), and 3D Skeleton information are proposed. The multimodal data features fusion helps to complete utilization of available sensor data and increased the recognition rate.
- It is observed that single modality based approaches are severely affected by self-occlusions, noise and error present in the video, especially in skeleton coordinates. Therefore, the RGB skeleton MHIs is utilized to eliminate the errors presenting in the estimation of 3D coordinates.
- Further, it is observed that fusion at the bottleneck layer before the top layer increases the classification accuracy with the SVM classifier as compared to softmax layer prediction scores.
- The order to test the efficacy of the proposed model we evaluated our model on four challenging RGB-D datasets and our model outperform the other state-of-the-art model. We achieved the highest accuracy on CAD-60, Florence 3D and SBU interaction dataset and comparable performance on UT Kinect dataset.

5.2 Future Prospective

The main objective of HAR system approaches is to automatically recognize activity in videos. Most of the handcrafted features solutions are developed and recognized an activity from action template building on a given set of videos. Instead of satisfactory results of many issues are unresolved till now. Due to the process of building the action template is more prone to human mistakes in handcrafted features descriptors. Therefore, we have to focus on new extensions to the latest model of action templates. We have to more conscious, especially on real-time applications of computer visions systems that iteratively apply the action templates on online streaming video.

It is observed that training deep ConvNet architectures requires a lot of labelled data to avoid the overfitting of the model. Further, the performance of action recognition architectures suffers due to inaccurate labels. However, it possible for a mixture of annotated and unannotated data for training the models. Therefore, it is required to design such ConvNet architectures for action recognition that can learn features from both labelled and unlabelled data.

- In future, a more realistic study may be conducted on the unconstrained dataset, and EMRFs can be used for many other applications such as visual sentiment representation and analysis, movie analysis, content-based recommender systems etc.
- In future work, depth frames along with 3D skeleton coordinate information and multi-view actions classes may be used to make the action prediction more dynamic to intra-class-life applications. It may also be applied for real-time detection of human activities, and other useful applications such as crowd anomaly detection, sports actions classification, and development of intelligent surveillance system, etc.

- However, multimodal features extraction from intermediate layers is more complicated due to a large number of deep layers of the pre-trained architectures. Therefore, we can try to develop such a deep model by utilizing the parallel computation and representation of CNN architecture. Hence, we can extend our approach for real-time applications, elderly care, and more complex actions with less complexity.

5.3 Future Applications

It is observed that most of the existing solutions based on handcrafted features or deep features are less suitable for real-time action recognition due to their dependencies on computational resources. It is necessary to develop such models that meet real-time action recognition in videos. In future, our motive to design and make necessary changes in the proposed frameworks to make them suitable for such challenges.

- **Real-time multi-person pose estimation**

Human pose estimation is the process of detecting the human pose in an image or video sequences. It is also called localization of human joints. A pose estimation task can be of any type such as single person pose estimation; Multiperson pose estimation and real-time pose estimation in public places. Real-time multi-person pose estimation is a challenging task as compared with others. It may include unknown number of persons at multiple time, orientations, scales, complex interactions, occlusions, and limb articulations. The complexities of the algorithm may grow with the number of peoples in the scene. Therefore, it is a challenging task and needs to explore for future research [211].

- **Suspicious Activity recognition**

In recent world video surveillance plays a major role in the security tasks [212]. The video surveillance system consisted of activities recognition or

abnormal activities detection and behavioural analysis that can be used for real-time applications. In future, we would like to extend our deep learning-based model for detection of abnormal or suspicious activities events in real-time during video surveillance.

- **Children Activity recognition**

Due to the prevalence of smartphones and advanced wireless technologies, the popularity of smartphones based activity recognition increasing daily, especially for mobile healthcare (mHealth). Till now, major activity recognition solutions focused on adult healthcare diseases such as asthma attacks [213]. However, recent studies and available physiological data impact of asthma exacerbation in children also [214]. In future, our target is developed activity model with the help of smart devices for prevention and recognition of asthma attacks in patients, especially in children. It seems to be more challenging because collecting large annotated database for children as compared with adult's database. Further, there is a large variation between children's activities when performing similar activities.

- **Autonomous Driving Vehicle**

Activity recognition is the most fundamental building block of autonomous driving vehicles. Action detection and prediction algorithms could be one of the potential application for self-driving vehicles. In future, we will try to extend our deep learning action classification model for real-time action prediction such as pedestrian's action detection.

References

- [1] S. Herath, M. Harandi and F. Porikli, "Going deeper into action recognition: A survey," *Image and Vision Computing*, vol. 60, pp. 4-21, 2017.
- [2] J. Liu, J. Luo and M. Shah, "Recognizing Realistic Actions from Videos "in the Wild"," in *IEEE International Conference on Computer Vision and Pattern Recognition(CVPR)*, 2009.
- [3] C. Dhiman and D. K. Vishwakarma, "A Robust Framework for Abnormal Human Action Recognition using R-Transform and Zernike Moments in Depth Videos," *IEEE Sensors Journal*, vol. 19, no. 13, pp. 5195-5203, 2019.
- [4] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia and A. Baskurt, "Sequential deep learning for human action recognition," in *Proceedings of the Second International Conference on Human Behavior Understanding*, 2011.
- [5] S. Ji, W. Xu, M. Yang and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221-231, 2013.
- [6] D. K. Vishwakarma, "Analysis of video sequence using intelligent techniques," Delhi Technological University, New Delhi, 2015.
- [7] C. Dhiman, "Identification Of Human Actions In Video Sequences," Delhi Technological University, New Delhi, 2019.
- [8] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. Int. Conference on Computer Vision (ICCV)*, 2013.
- [9] H. Wang, A. Klaeser, C. Schmid and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *IJCV*, 2013.
- [10] L. Shao, X. Zhen, D. Tao and X. Li, "Spatio-temporal Laplacian pyramid coding for action recognition," *IEEE Trans. Cybern.*, vol. 44, no. 6, p. 817-827, 2014.

- [11] D. K. Vishwakarma and T. Singh, "A visual cognizance based multi-resolution descriptor for human action recognition using key pose," *AEU - International Journal of Electronics and Communications*, vol. 107, pp. 157-169, 2019.
- [12] D. K. Vishwakarma and K. Singh, "Human Activity Recognition based on Spatial Distribution of Gradients at Sub-levels of Average Energy Silhouette Images," *IEEE Transactions on Cognitive and Development Systems*, vol. 99, pp. 1-1, 2016.
- [13] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *CVPR*, 2014.
- [14] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.
- [15] J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation," in *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [16] D. Michelsanti, Z.-H. Tan, S. Sigurdsson and J. Jensen, "Deep-learning-based audio-visual speech enhancement in presence of Lombard effect," *Speech Communication*, vol. 115, pp. 38-50, 2019.
- [17] Z. Teng, B. Zhang and J. Fan, "Three-step action search networks with deep Q-learning for real-time object tracking," *Pattern Recognition*, vol. 101, p. 107188, 2020.
- [18] C. Dong, C. C. Loy, K. He and X. Tang, "Learning a deep convolutional network for image super-resolution," in *ECCV*, 2014.
- [19] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Survey*, vol. 43, no. 3, pp. 1-43, 2011.
- [20] J. Aggarwal and L. Xia, "Human activity recognition from 3D data- A review," *Pattern Recognition Letters*, vol. 48, 2013.
- [21] T. Singh and D. K. Vishwakarma, "Video benchmarks of human action datasets: a review," *Artificial Intelligence Review*, vol. 52, no. 2, p. 1107-1154, 2018.

- [22] S. Vishwakarma and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance," *The Visual Computer*, vol. 29, no. 10, p. 983–1009, 2013.
- [23] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Advances in Neural Information Processing Systems*, 2014.
- [24] D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. Int. Conference on Computer Vision*, 2015.
- [25] X. S. Nguyen, A. I. Mouaddib and T. P. Nguyen, "Hierarchical Gaussian descriptor based on local pooling for action recognition," *Machine Vision and Applications*, vol. 30, pp. 321-343, 2019.
- [26] X. Ji, J. Cheng, W. Feng and D. Tao, "Skeleton embedded motion body partition for human action recognition using depth sequences," *Signal Processing*, vol. 143, pp. 56-68, 2017.
- [27] Y. Ji, Y. Yang, X. Xu and H. T. Shen, "One-shot learning based pattern transition map for action early recognition," *Signal Processing*, vol. 143, pp. 364-370, 2018.
- [28] P. Khaire, P. Kumar and J. Imran, "Combining CNN streams of RGB-D and skeletal data for human activity recognition," *Pattern Recognition Letters*, vol. 000, pp. 1-10, 2018.
- [29] C. Schuldt, I. Laptev and B. Caputo, "Recognizing human actions: a local SVM approach," in *17th International Conference on Pattern Recognition*, 2004.
- [30] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257-267, 2001.
- [31] F. C. Heilbron, V. Escorcia, B. Ghanem and J. C. Niebles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 2015.
- [32] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, 2014.

- [33] Z. Liu, L. Zhou, H. Leung and H. P. H. Shum, "Kinect Posture Reconstruction Based on a Local Mixture of Gaussian Process Models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 11, pp. 2437-2450, 2016.
- [34] L. Fu, J. Zhang and K. Huang, "ORGM: Occlusion Relational Graphical Model for Human Pose Estimation," *IEEE Transaction on Image Processing*, vol. 26, no. 2, pp. 927-941, 2017.
- [35] T. Hao, D. Wu, Q. Wang and J.-S. Sun, "Multi-view representation learning for multi-view action recognition," *Journal of Visual Communication and Image Representation*, 2017.
- [36] C. Feichtenhofer, A. Pinz and R. P. Wildes, "Spatiotemporal Multiplier Networks for Video Action Recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, hawaii, 2017.
- [37] P. Foggia, A. Saggese, N. Strisciuglio and M. Vento, "Exploiting the deep learning paradigm for recognizing human actions," in *IEEE AVSS*, 2014.
- [38] S. Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan and S. Vijayanarasimhan, "YouTube-8M: {A} Large-Scale Video Classification Benchmark," in *CoRR*, 2016.
- [39] H. Zhang, W. Zhou and L. E. Parker, "Fuzzy Temporal Segmentation and Probabilistic Recognition of Continuous Human Daily Activities," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 5, pp. 598-611, 2015.
- [40] A. Pentland, "Smart rooms, smart clothes," in *Fourteenth Int. Conf. on Pattern Recognition*, 1998.
- [41] S. Sarkar, P. Phillips, Z. Liu, I. Vega, P. Grother and K. Bowyer, "The humanoid gait challenge problem: datasets, performance, and analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 162-177, 2005.
- [42] S. Chang, "The holy grail of content-based media analysis," *IEEE Multimedia*, vol. 9, no. 2, pp. 6-10, 2002.
- [43] Y. Rui and T. Huang, "Image retrieval: current techniques, promising directions and open issues," *J. Vis. Commun. Image Represent.*, vol. 10, p. 39-62, 1999.

- [44] W. Hu, D. Xie, Z. Z. Fu and S. Maybank, " Semantic-based surveillance video retrieval," *IEEE Transactions on Image Processing* , vol. 16, no. 4, p. 1168–1181, 2007.
- [45] D. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien and D. Ramanan, "Computational studies of human motion: part 1, tracking and motion synthesis.," *Found. Trends Comput. Graph. Vis.* , pp. 77-254, 2005.
- [46] C. Stauffer and W. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, p. 22, 2000.
- [47] S. Singh, S. A. Velastin and H. Ragheb, "Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods," in *Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2010.
- [48] B. B. Amor, J. Su and A. Srivastava, "Action Recognition Using Rate-Invariant Analysis of Skeletal Shape Trajectories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 1-13, 2016.
- [49] J. Feng, S. Zhang and J. Xiao, "Explorations of skeleton features for LSTM-based action recognition," *Multimed Tools and Application*, pp. 1-13, 2017.
- [50] D. Hogg, "Model-based vision: a program to see a walking person," *Image and Vision Computing*, vol. 1, no. 1, p. 5–20, 1983.
- [51] K. Rohr, "Towards Model-Based Recognition of Human Movements in Image Sequences," *CVGIP: Image Understanding* , vol. 59, no. 1, p. 94–115, 1994.
- [52] Y. Tian, L. Cao, Z. Liu and Z. Zhang, " Hierarchical filtered motion for action recogniion in crowded videos," *IEEE Trans. Syst. Man Cybern*, vol. 42, no. 3, pp. 313-323, 2012.
- [53] C. Harris and M. Stephens, "A combined corner and edge detector," in *Fourth Alvey Vision Conference*, 1988.
- [54] M. Blank, L. Gorelick, E. Shechtman, M. Irani and R. Basri, "Actions as space-time shapes," in *Tenth IEEE International Conference on Computer Vision (ICCV'05)*, Beijing,, 2005.

- [55] D. Weinland, R. Ronfard and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2-3, pp. 249-257, 2006.
- [56] E. Shechtman and M. Irani, "Space-time behaviour based correlation," in *IEEE Conference on Computer Vision and Pattern Analysis*, Los Alamitos, CA, 2005.
- [57] M. D. Rodriguez, J. Ahmed and M. Shah, "Action MACH: A Spatio-temporal Maximum Average Correlation Height Filter for Action Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, 2008.
- [58] A. Yilmaz and M. Shah, "Actions sketch: a novel action representation," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005.
- [59] I. Laptev, "On Space-Time Interest Points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107-123, 2005.
- [60] G. Willems, T. Tuytelaars and L. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proc. European Conference on Computer Vision (ECCV)*, 2008.
- [61] P. Dollar, V. Rabaud, G. Cottrell and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005.
- [62] A. Kläser, M. Marszałek and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *British Machine Vision Conference*, Leeds, 2008.
- [63] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*, 2001.
- [64] I. Laptev, M. Marszałek, C. Schmid and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, 2008.
- [65] N. Dalal, B. Triggs and C. Schmid, "Human Detection Using Oriented Histograms of Flow and Appearance," in *Lecture Notes in Computer Science*, 428-441, 2006.

- [66] V. Kantorov and I. Laptev, "Efficient feature extraction, encoding, and classification for action recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
- [67] P. Matikainen, M. Hebert and R. Sukthankar, "Trajectons: action recognition through the motion analysis of tracked features," in *IEEE 12th International Conference on Computer Vision*, 2009.
- [68] R. Messing, C. Pal and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *Int. Conference on Computer Vision (ICCV)*, 2009.
- [69] B. D. Kanade. and T. Lucas, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, 1981.
- [70] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu and C.-W. Ngo, "Trajectory-based modeling of human actions with motion reference points," in *Proc. European Conference on Computer Vision (ECCV)*, 2012.
- [71] H. Wang, M. Ullah, A. Kläser, I. Laptev and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *British Machine Vision Conference*, 2009.
- [72] L. Brun, G. Percannella, A. Saggesea and M. Vento, "Action recognition by using kernels on aclets sequences," *Computer Vision and Image Understanding*, vol. 144, pp. 3-13, 2016.
- [73] Y. Zhu, X. Zhao, Y. Fu and Y. Liu, "Sparse coding on local spatial-temporal volumes for human action recognition," in *Proc. Asian Conference on Computer Vision*, 2011.
- [74] G. Csurka, C. Dance, L. Fan, J. Willamowski and C. Bray, "Visual categorization with bags of keypoints," in *In Workshop on Statistical Learning in Computer Vision*, 2004.
- [75] T. Guha and R. K. Ward, "Learning sparse representations for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1576-1588, 2012.
- [76] G. Somasundaram, A. Cherian, V. Morellas and N. Papanikolopoulos, "Action recognition using global spatio-temporal features derived from sparse representations," *Comput. Vis. Image Underst.*, vol. 123, pp. 1-13, 2014.

- [77] S. Sadanand and J. Corso, "Action Bank: A High-Level Representation of Activity in Video," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [78] A. Gaidon, Z. Harchaoui and C. Schmid, "Actom sequence models for efficient action detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
- [79] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Proc. IEEE* 77 (2) , 1989.
- [80] S. Hongeng and R. Nevatia, "Large-scale event detection using semi-hidden Markov models," in *Proc. Int. Conference on Computer Vision (ICCV)*, 2003.
- [81] K. Tang, L. F. Fei and D. Koller, "Learning latent temporal structure for complex event detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [82] C. Sun and R. Nevatia, "ACTIVE: activity concept transitions in video event classification," in *Proc. Int. Conference on Computer Vision (ICCV)*, 2013.
- [83] V. Carletti, P. Foggia, G. Percannella, A. Saggese and M. Vento, "Recognition of human actions from RGB-D videos using a reject option," in *International Workshop on Social Behaviour Analysis* , 2013.
- [84] Y. Wang and G. Mori, "Hidden part models for human action recognition: probabilistic versus max margin," *IEEE Trans. Pattern Anal. Mach. Intell.* , vol. 33, no. 7, pp. 1310-1323, 2011.
- [85] A.-A. Liu, Y.-T. Su, W.-Z. Nie and M. Kankanhalli, "Hierarchical Clustering Multi-Task Learning for Joint Human Action Grouping and Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 1, pp. 102-114, 2017.
- [86] W. Xu, Z. Miao, X. P. Zhang and Y. Tian, "A Hierarchical Spatio-Temporal Model for Human Activity Recognition," *IEEE Transactions on Multimedia*, vol. 99, pp. 1-1, 2017.
- [87] Y. Shan, Z. Zhang, P. Yang and K. Huang, "Adaptive Slice Representation for Human Action Classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 10, pp. 1624-1636, 2015.

- [88] Y. Yu, J. Choi, Y. Kim, K. Yoo, S.-H. Lee and G. Kim, "Supervising Neural Attention Models for Video Captioning by Human Gaze Data," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Hawaii, 2017.
- [89] T. H. The, B.-V. Le, S. Lee and Y. Yoon, "Interactive activity recognition using pose-based spatio-temporal relation features and four-level Pachinko Allocation Model," *Informatics and Computer Science Intelligent Systems Applications*, vol. 369, p. 317-333, 2016.
- [90] H. Yan, "Discriminative sparse projections for activity-based person recognition," *Neurocomputing*, vol. 208, p. 183-192, 2016.
- [91] J. Yuan, B. Ni, X. Yang and A. A. Kassim, "Temporal Action Localization with Pyramid of Score Distribution Features," in *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016.
- [92] Z. Wang, L. Wang, W. Du and Y. Qiao, "Exploring Fisher Vector and Deep Networks for Action Spotting," in *CVPR*, 2015.
- [93] D. K. Vishwakarma, R. Kapoor and A. Dhiman, "A proposed framework for the recognition of human activity by exploiting the characteristics of action dynamics," *Robotics and Autonomous System*, vol. 77, pp. 25-38, 2016.
- [94] S. Agahian, F. Negin and C. Köse, "Improving bag-of-poses with semi-temporal pose descriptors for skeleton-based action recognition," *The Visual Computer*, pp. 1-17, 2018.
- [95] X. Peng, C. Zou, Y. Qiao and Peng.Q., "Action recognition with stacked fisher vectors," in *ECCV*, 2014.
- [96] W. Yang, Y. Wang and G. Mori, "Recognizing human actions from still images with latent poses," in *CVPR*, 2010.
- [97] V. Delaitre, I. Laptev and J. Sivic, "Recognizing human actions in still images: a study of bag-of-features and part-based representations," in *Proceedings of the British Machine Vision Conference*, 2010.
- [98] Y. Zhang, L. Cheng, J. Wu, J. Cai, M. N. Do and J. Lu, "Action Recognition in Still Images With Minimum Annotation Efforts," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5479-5490, 2016.

- [99] C. Thureau and V. Hlavac, "Pose primitive based human action recognition in videos or still images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.
- [100] K. Raja, I. Laptev, P. Pérez and L. Oisel, "Joint pose estimation and action recognition in image graphs," in *IEEE International Conference on Image Processing*, Brussels, 2011.
- [101] N. Guan, D. Tao, L. Lan, Z. Luo and X. Yang, "Activity Recognition from Still Images with Transductive Non-negative Matrix Factorization," in *ECCV*, 2014.
- [102] K. Zhao, A. Alavi, A. Wiliem and C. Lovell, "Efficient clustering on Riemannian manifolds : A kernelised random projection approach," *Pattern Recognition*, vol. 51, pp. 333-345, 2015.
- [103] A. Saggese, N. Strisciuglio, M. Vento and N. Petkov, "Learning skeleton representations for human action recognition," *Pattern Recognition Letters*, pp. 1-9, 2018.
- [104] I. Laptev and T. Lindeberg, "Local Descriptors for Spatio-Temporal Recognition," in *ECCV Workshop on Spatial Coherence for Visual Motion Analysis*, 2004.
- [105] M. Al-Nawashi, O. Al-Hazaimeh and M. Saraee, "A novel framework for intelligent surveillance system based on abnormal human activity detection in academic environments," *Neural Computing and Applications*, vol. 28, p. 565-572, 2017.
- [106] S. Ghodsi, H. Mohammadzade and E. Korki, "Simultaneous joint and object trajectory templates for human activity recognition from 3-D data," *J. Vis. Commun. Image Ranging*, vol. 55, pp. 729-741, 2018.
- [107] Y. Yang, C. Deng, D. Tao, S. Zhang, W. Liu and X. Gao, "Latent Max-Margin Multitask Learning With Skelets for 3-D Action Recognition," *IEEE Transactions on Cybernetics*, vol. 47, no. 2, pp. 439-448, 2017.
- [108] E. Shabaninia, A. R. N. Nilchi and S. Kasaei, "A weighting scheme for mining key skeletal joints for human action recognition," *Multimedia Tools and Applications*, vol. 78, p. 31319-31345, 2019.
- [109] J. Kong, T. Liu and M. Jiang, "Collaborative multimodal feature learning for RGB-D action recognition," *J. Vis. Commun. Image Ranging*, vol. 59, pp. 537-549, 2019.

- [110] N. Raman and S. Maybank, "Activity recognition using a supervised non-parametric hierarchical HMM," *Neurocomputing*, vol. 19, p. 163–177, 2016.
- [111] X. Yan, H. Chang, S. Shan and X. Chen, "Modeling video dynamics with deep dynencoder," in *European Conference on Computer Vision*, 2014.
- [112] J.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga and G. Toderici, "Beyond short snippets: deep networks for video classification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [113] J. Donahue, L. Hendricks, S. Guadarrama, M. V. Rohrbach, K. Saenko and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [114] Y. Han, P. Zhang, T. Zhuo, W. Huang and Y. Zhang, "Going deeper with two-stream ConvNets for action recognition in video surveillance," *Pattern Recognition Letters*, 2017.
- [115] M. Safaei and H. Foroosh, "Single Image Action Recognition by Predicting Space-Time Saliency," *arXiv:1705.04641v1 [cs.CV]*, pp. 1-9, 2017.
- [116] M. Liu, H. Liu and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognition*, vol. 68, pp. 346-361, 2017.
- [117] C. Feichtenhofer, A. Pinz and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. pp. 1933–1941, 2016.
- [118] A. S. Keçeli, A. Kaya and A. B. Can, "Combining 2D and 3D deep models for action recognition with depth information," *Signal, Image and Video Processing*, vol. 12, pp. 1197-1205, 2018.
- [119] E. P. Ijjina and K. M. Chalavadi, "Human action recognition in RGB-D videos using motion sequence information and deep learning," *Pattern Recognition*, vol. 72, p. 504–516, 2017.
- [120] C. Jing, P. Wei, H. Sun and N. Zheng, "Spatiotemporal neural networks for action recognition based on joint loss," *Neural Computing and Applications*, 2019.

- [121] D. Srihari, P. V. V. · Kishore, E. K. Kumar, A. Kumar, M. T. K. Kumar, M. V. D. Prasad and C. R. Prasad, "A four-stream ConvNet based on spatial and depth flow for human action classification using RGB-D data," *Multimedia Tools and Applications*, 2020.
- [122] A. Elboushaki, R. Hannane, K. Afdel and L. Koutti, "MultiD-CNN: A multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image," *Expert Systems With Applications*, vol. 139, 2020.
- [123] N. Srivastava, E. Mansimov and R. Salakhutdinov, " Unsupervised Learning of Video Representations Using LSTMs," *CoRR*, 2015.
- [124] I. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, 2014.
- [125] M. Mathieu, C. Couprie and Y. LeCun, " Deep Multi-Scale Video Prediction Beyond Mean Square Error," *CoRR*, 2015.
- [126] L. Wang, Y. Xiong, D. Lin and L. Van Gool, "UntrimmedNets for Weakly Supervised Action Recognition and Detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Hawaii, 2017.
- [127] A. Cherian, B. Fernando, M. Harandi and S. Gould, "Generalized Rank Pooling for Activity Recognition," in *CVPR*, Hawaii, 2017.
- [128] I. Misra, C. Zitnick and M. Hebert, " Unsupervised Learning Using Sequential Verification for Action Recognition," *arXiv preprint arXiv:1603.08561*, 2016.
- [129] C. Lea, M. D. Flynn, R. Vidal, A. Reiter and G. D. Hager, "Temporal Convolutional Networks for Action Segmentation and Detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Hawaii, 2017.
- [130] B. Fernando, E. Gavves, M. Oramas, A. Ghodrati and T. Tuytelaars, "Modeling video evolution for action recognition," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [131] B. Fernando and S. Gould, "Learning end-to-end video classification with rank-pooling," in *ICML*, 2016.

- [132] B. Ni, P. Moulin, X. Yang and S. Yan, "Motion Part Regularization: Improving action recognition via trajectory group selection," in *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, 2015.
- [133] I. Abbasnejad, S. Sridharan, S. Denman, C. Fookes and S. Lucey, "Complex Event Detection Using Joint Max Margin and Semantic Features," in *International Conference on Digital Image Computing: Techniques and Applications*, Gold Coast, 2016.
- [134] S. Nazir, M. H. Yousaf, J.-C. Nebel and S. A. Velastin, "A Bag of Expression framework for improved human action recognition," *Pattern Recognition Letters*, vol. 103, p. 39–45, 2018.
- [135] Y. J. Kim, N. G. Cho and S. W. Lee, "Group Activity Recognition with Group Interaction Zone," in *22nd International Conference on Pattern Recognition*, Stockholm, 2014.
- [136] I. C. Duta, B. Ionescu, K. Aizawa and N. Sebe, "Spatio-Temporal Vector of Locally Max Pooled Features for Action Recognition in Videos," in *CVPR*, Hawaii, 2017.
- [137] Y. G. Jiang, Z. Wu, J. Wang, X. Xue and S. F. Chang, "Exploiting Feature and Class Relationships in Video Categorization with Regularized Deep Neural Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, pp. 1-1, 2017.
- [138] B. Singh, T. Marks, M. Jones and C. Tuzel, "A Multi-Stream Bi-Directional Recurrent Neural Network for Fine-Grained Action Detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [139] A. Miech, I. Laptev and J. Sivic, "Learnable pooling with Context Gating for video classification," in *CVPR Workshop*, Hawaii, 2017.
- [140] M. Barekatin and e. al., "Okutama-Action: An Aerial View Video Dataset for Concurrent Human Action Detection," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Honolulu, 2017.
- [141] R. Goyal, S. E. Kahou, V. Michalski, J. Materzyńska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. M. Freitag, F. Hoppe, C. Thureau, I. Bax and R. Memisevic, "The "something something" video database for learning and evaluating visual common sense," *arXiv:1706.04261v2 [cs.CV]*, 2018.

- [142] C. Yang, Y. Xu, J. Shi, B. Dai and B. Zhou, "Temporal Pyramid Network for Action Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [143] L. Chen, H. Wei and J. Ferryman, "ReadingAct RGB-D action dataset and human action recognition from local features," *Pattern Recognition Letters*, vol. 50, p. 159–169, 2014.
- [144] A. Franco, A. Magnani and D. Maio, "A multimodal approach for human activity recognition based on skeleton and RGB data," *Pattern Recognition Letters*, vol. 131, pp. 293–299, 2020.
- [145] R. Chaudhry, F. Ofli, G. Kurillo, R. Bajcsy and R. Vidal, "Bio-inspired Dynamic 3D Discriminative Skeletal Features for Human Action Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Portland, 2013.
- [146] J. Tayyub, A. Tavanai, Y. Gatsoulis, A. Cohn and D. Hogg, "Qualitative and quantitative spatio-temporal relations," in *ACCV*, 2015.
- [147] A. Iosifidis, A. Tefas, N. Nikolaidis and I. Pitas, "Human action recognition in stereoscopic videos based on a bag of features and disparity pyramids," in *22nd European Signal Processing Conference*, Lisbon, 2014.
- [148] T. Singh and D. K. Vishwakarma, "A Deeply Coupled ConvNet for Human Activity Recognition using Dynamic and RGB Images," *Neural Computing and Applications*, 2020.
- [149] C. Chen, R. Jafari and N. Kehtarnavaz, "UTD-MHAD: A Multimodal Dataset for Human Action Recognition Utilizing a Depth Camera and a Wearable Inertial Sensor," in *Proceedings of IEEE International Conference on Image Processing*, Canada, 2015.
- [150] C. Wolf, J. Mille, E. Lombardi, O. Celiktutan, M. Jiu, E. Dogan, G. Eren, M. Baccouche, E. Dellandrea, C.-E. Bichot, C. Garcia and B. Sankur, "Evaluation of video activity localizations integrating quality and quantity measurements," *Computer Vision and Image Understanding*, vol. 127, pp. 14–30, 2014.
- [151] A. Jalal and Y. Kim, "Dense Depth Maps-based Human Pose Tracking and Recognition in Dynamic Scenes Using Ridge Data," in *11th IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2014.

- [152] W. Nie, W. Wang and X. Huang, "SRNet: Structured Relevance Feature Learning Network From Skeleton Data for Human Action Recognition," *IEEE Access*, vol. 7, pp. 132161-132172, 2019.
- [153] G. Vaquette, A. L. Orcesi and C. Achard, "The DAily Home LIfe Activity Dataset: A High Semantic Activity Dataset for Online Recognition," in *IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, Washington, DC, 2017.
- [154] F. Negin, P. Rodriguez, M. Koperski, A. Kerboua, J. Gonzalez, J. Bourgeois, E. Chapoulie, P. Robert and F. Bremond, "PRAXIS: Towards Automatic Cognitive Assessment Using Gesture Recognition," *Expert Systems with Applications*, vol. 106, pp. 21-35, 2018.
- [155] P. Zeng and Z. Chen, "Perceptual quality measure using JND model of the human visual system," in *IEEE International Conference on Electric Information and Control Engineering*, 2011.
- [156] L. A. Zadeh, "Fuzzy sets," *Information Control*, vol. 8, no. 3, pp. 338-353, 1965.
- [157] J. Canny, "A Computational Approach To Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, p. 679-698, 1986.
- [158] S. Arivazhagan, L. Ganesan and S. P. Priyal, "Texture classification using Gabor wavelets based rotation invariant features," *Pattern Recognition Letters*, pp. 1976-1982, 2006.
- [159] A. Pasupathy, Y. El-Shamayleh and D. V. Popovkina, "Visual Shape and Object Perception," *Neuroscience*, 2018.
- [160] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008.
- [161] V. Vapnik, "An overview of statistical learning theory," *IEEE Transaction Neural Network*, vol. 10, no. 5, pp. 989-99, 1999.
- [162] T. Cover and P. Hart, "Nearest neighbour pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, p. 1967, 21-27.
- [163] Y. Zheng, H. Yao, X. Sun, S. Zhao and F. Porikli, "Distinctive action sketch for human action recognition," *Signal Processing*, vol. 144, p. 323-332, 2018.

- [164] B. Saghafi and D. Rajan, "Human action recognition using Pose-based discriminant embedding," *Signal Processing: Image Communication*, vol. 27, pp. 96-111, 2012.
- [165] A. Iosifidis, A. Tefas and I. Pitas, "Discriminant bag of words based representation for human action recognition," *Pattern Recognition Letters*, vol. 49, pp. 185-192, 2014.
- [166] D. K. Vishwakarma and R. Kapoor, "Hybrid classifier based human action recognition using silhouettes and cells," *Expert Systems with Applications*, vol. 42, no. 20, pp. 6957-6965, 2015.
- [167] B. Wang, Y. Hu, J. Gao, Y. Sun and B. Yin, "Localized LRR on Grassmann Manifold: An Extrinsic View," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.
- [168] A. A. Chaaraoui, P. C. Pérez and F. F. Revuelta, "Silhouette-based human action recognition using sequences of key poses," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1799-1807, 2013.
- [169] J. C. Niebles and L. Fei-Fei, "A Hierarchical Model of Shape and Appearance for Human Action Classification," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007.
- [170] C. Thureau, "Behavior Histograms for Action Recognition and Human Detection," *Lecture Notes in Computer Science, Springer*, vol. 4814, pp. 299-312, 2007.
- [171] A. Eweiwia, M. S. Cheema and C. Bauckhage, "Action recognition in still images by learning spatial interest regions from videos," *Pattern Recognition Letters*, vol. 51, pp. 8-15, 2015.
- [172] S. Baysal and P. Duygulu, "A line based pose representation for human action recognition," *Signal processing: Image Communication*, vol. 28, pp. 458-471, 2013.
- [173] G. Batchuluun, J. H. Kim, H. G. Hong, J. K. Kang and K. R. Park, "Fuzzy system based human behavior recognition by combining behavior prediction and recognition," *Expert Systems With Applications*, vol. 81, p. 108-133, 2017.
- [174] Y. Wang and G. Mori, "Human Action Recognition Using Semi-Latent Topic Model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1762-1764, 2009.
- [175] N. Ikizler-Cinbis and S. Sclaroff, "Object, scene and actions: combining multiple features for human action recognition," in *Computer Vision-ECCV*, 2010.


- [176] Q. Le, W. Zou, S. Yeung and A. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.
- [177] Y. Yi and M. Lin, "Human action recognition with graph-based multiple-instance learning," *Pattern Recognition*, vol. 53, pp. 148-162, 2016.
- [178] L. Shao, L. Liu and M. Yu, "Kernelized multiview projection for robust action recognition," *Int. J. Comput. Vis.*, pp. 1-15, 2015.
- [179] H.-J. Jung and K.-S. Hong, "Modeling temporal structure of complex actions using Bag-of-Sequencelets," *Pattern Recognition Letters*, vol. 85, pp. 21-28, 2017.
- [180] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [181] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 1997, pp. 1735-1780, 1997.
- [182] R. J. Williams, G. E. Hinton and D. E. Rumelhart, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, p. 533-536, 1986.
- [183] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi and S. Gould, "Dynamic Image Networks for Action Recognition," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 3034-3042., 2016.
- [184] A. J. Smola and B. Scholkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, p. 199-222, 2004.
- [185] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg and D. Samaras, "Two-person Interaction Detection Using Body-Pose Features and Multiple Instance Learning," in *IEEE International Conference Computer Vision and Pattern Recognition Workshops (CVPRW)*, Rhode Island, 2012.
- [186] J. Wang, Z. Liu, Y. Wu and J. Yuan, "Mining Actionlet Ensemble for Action Recognition with Depth Cameras," in *IEEE Conference on Computer Vision and Pattern Recognition*, Rhode Island, 2012.

- [187] O. Oreifej and Z. Liu, "HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Portland, Oregon, 2013.
- [188] P. D. Kingma and J. L. Ba, "ADAM: A Method for Stochastic Optimization," in *International Conference on Learning Representations*, San Diego, 2015.
- [189] H. B. Mann and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50-60, 1947.
- [190] C. Jia, Y. Kong, Z. Ding and Y. Fu, "Latent Tensor Transfer Learning for RGB-D Action Recognition," in *Proceedings of the 22nd ACM international conference on Multimedia*, Orlando, Florida, USA, 2014.
- [191] R. Vemulapalli and R. Chellapa, "Rolling Rotations for Recognizing Human Actions From 3D Skeletal Data," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [192] L. Seidenari, V. Varano, S. Berretti, A. D. Bimbo and P. Pala, "Recognizing Actions from Depth Cameras as Weakly Aligned Multi-part Bag-of-Poses," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Portland, 2013.
- [193] X. Cai, W. Zhou, L. Wu, J. Luo and H. Li, "Effective Active Skeleton Representation for Low Latency Human Action Recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 141-154, 2016.
- [194] H. Zhang and L. E. Parker, "Bio-inspired predictive orientation decomposition of skeleton trajectories for real-time human activity prediction," in *IEEE International Conference on Robotics and Automation (ICRA)*, Seattle, WA, 2015.
- [195] T.-T. Huynh, C.-H. Hua, N. A. Tu, T. Hur, J. Bang, D. Kim, M. B. Amin, B. H. Kang, H. Seung, S.-Y. Shin, E.-S. Kim and S. Lee, "Hierarchical topic modeling with pose-transition feature for action recognition using 3D skeleton data," *Information Sciences*, vol. 444, pp. 20-35, 2018.
- [196] C. Szegedy, S. Ioffe, V. Vanhoucke and A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," *arXiv:1602.07261v2 [cs.CV]*, 2016.

- [197] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.
- [198] C. N. Phyto, T. T. Zin and P. Tin, "Deep Learning for Recognizing Human Activities Using Motions of Skeletal Joints," *IEEE Transaction on Consumer Electronics*, vol. 65, no. 2, pp. 243-252, 2019.
- [199] M. Haghighat, M. A. Mottaleb and W. Alhalabi, "Discriminant Correlation Analysis: Real-Time Feature Level Fusion for Multimodal Biometric Recognition," *IEEE Transactions On Information Forensics and Security*, vol. 11, no. 9, pp. 1984-1996, 2016.
- [200] L. Xia, C. Chen and J. Aggarwal, "View Invariant Human Action Recognition Using Histograms of 3D Joints," in *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012.
- [201] J. Sung, C. Ponce, B. Selman and A. Saxena, "Unstructured human activity detection from RGBD images," in *IEEE International Conference on Robotics and Automation*, Saint Paul, MN, 2012.
- [202] D. Avola, M. Bernardi and G. L. Foresti, "Fusing depth and colour information for human action recognition," *Multimedia Tools and Applications*, vol. 78, p. 5919-5939, 2019.
- [203] M. Devanne, H. Wannous, S. Berretti, P. D. Pala and A. D. Bimbo, "3-D Human Action Recognition by Shape Analysis of Motion Trajectories on Riemannian Manifold," *IEEE Transactions on Cybernetics*, vol. 45, no. 7, pp. 1340-1352, 2015.
- [204] S. Zhang, X. Liu and J. Xiao, "On Geometric Features for Skeleton-Based Action Recognition Using Multilayer LSTM Networks," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, Santa Rosa, CA, 2017.
- [205] R. Vemulapalli, F. Arrate and R. Chellappa, "Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group," in *IEEE Conference on CVPR*, Columbus, OH, 2014.
- [206] T. Liu, J. Wang, S. Hutchinson and M. Q.-H. Meng, "Skeleton-Based Human Action Recognition by Pose Specificity and Weighted Voting," *International Journal of Social Robotics*, vol. 11, p. 219-234, 2019.

- [207] Q. Li, W. Lin and J. Li, "Human activity recognition using dynamic representation and matching of skeleton feature sequences from RGB-D images," *Signal Processing: Image Communication*, vol. 68, pp. 265-272, 2018.
- [208] A. A. A. Salih and C. Youssef, "Spatiotemporal representation of 3D skeleton joints-based action recognition using modified spherical harmonics," *Pattern Recognition Letters*, vol. 83, pp. 32-41, 2016.
- [209] B. Sun, D. Kong, S. Wang, L. Wang and Y. Y. Wang, "Effective human action recognition using global and local offsets of skeleton joints," *Multimed Tools and Applications*, vol. 78, pp. 6329-6353, 2019.
- [210] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition.," *arXiv:arXiv:1409.1556*, 2014.
- [211] Z. Cao, T. Simon, S. E. Wei and Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," *CoRR*, 2016.
- [212] U. M. Kamthe and C. G. Patil, "Suspicious Activity Recognition in Video Surveillance System," in *Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, Pune, India, 2018.
- [213] A. Hosseini, C. M. Buonocore, S. Hashemzadeh, H. Hojaiji, H. Kalantarian, C. Sideris, A. A. Bui, C. E. King and M. Sarrafzadeh, "Feasibility of a secure wireless sensing smartwatch application for the self-management of pediatric asthma," *Sensors*, vol. 8, p. 18, 2017.
- [214] D. K. Mitchell, S. J. Kopel, C. A. Esteban, R. Seifer, N. W. Vehse, S. Chau and E. Jelalian, "Asthma status and physical activity in urban children," *Novel Epidemiology, Management, And Outcomes In Asthma*, 2017.

Author Biography

	<p>TEJ SINGH</p> <p>(2K16/Ph.D./EC/05)</p> <p>Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India</p>
---	--

Tej Singh received B.Tech. Degree from Madan Mohan Malaviya University of Technology, Gorakhpur, India, in the year 2010 and the M.E degree from the Thapar Institute of Engineering and Technology, Patiala, India, in the year 2014. Currently, he is working toward the fulfilment of PhD degree in the Department of Electronics and Communication Engineering, Delhi Technological University, New Delhi, India. His current research interests include image processing, pattern analysis and machine learning and deep learning human, artificial intelligence action and activity recognition.