IDENTIFICATION OF HUMAN ACTIONS IN VIDEO SEQUENCES

A Thesis Submitted to

Delhi Technological University

For the Award of Degree of

Doctor of Philosophy

In

Electronics and Communication Engineering

By

CHHAVI DHIMAN

(2K16/Ph.D./EC/07)

Under the Supervision of

Dr. Dinesh Kumar Vishwakarma

Associate Professor, Department of Information Technology

Department of Electronics & Communication Engineering
Delhi Technological University
(Formerly Delhi college of Engineering)
Delhi-110042, India
NOVEMBER-2019

DECLARATION

I declare that the research work reported in the thesis entitled **“Identification of Human Actions in 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:

(Chhavi Dhiman)
2K16/Ph.D./EC/07
Department of Electronics and Communication Engineering
Delhi Technological University,
Delhi-110042, India

CERTIFICATE

This is to certify that the thesis entitled **"Identification of Human Actions in Video Sequences"**, submitted by Ms. **Chhavi Dhiman** (Reg. No.: 2K16/PHD/EC/07) for the award of degree of Doctor of Philosophy, is based on the original research work carried out by her, to the Delhi Technological University. She has worked under my supervision and has fulfilled the requirements, which to our knowledge have reached the requisite standard for the submission of this thesis. It is further certified that the work embodied in this thesis is 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

ACKNOWLEDGMENTS

I owe tremendous debt and would like to express deep feelings of gratitude for the support and guidance of several people who have helped me to accomplish the research program with the support and direction of several persons. This challenging and rewarding experience has definitely helped me grow in character as well as academically. It gives me a great pleasure to now have the opportunity to express my gratitude towards them.

First and foremost, I wish to express my deep sense of gratitude and appreciation to my supervisor **Dr. Dinesh Kumar Vishwakarma**, for his stimulating supervision with invaluable suggestions, keen interest, constructive criticisms and constant encouragement during the course of my research work. I learned a lot from him not only in the field of academic but in other spheres of life also. I wholeheartedly acknowledge his full cooperation that I received from the very beginning of this work up to the completion in the form of this thesis and his guidance and support throughout the course of my Ph.D. research. His substantial and thorough approach, together with his genuine interest in the research, turned my research work into a great experience.

I would like to make a special note of thanks to all the research scholars in Biometric Research Laboratory, and Mrs. Sushma Vishwakarma and Little Diya for their co-operation and making the work environment so positive and vibrant with their presence.

Finally, I express my heartfelt gratitude to my highly respectable and adorable father and mother for their unconditional love, encouragement and blessings. They have been a guiding force all their life and tried to measure up to their expectations. I also express my abounding feelings of gratitude to my loving son, dear husband and respectable mother-in-law and father-in-law for the beautiful partnership of my responsibilities that we share and supporting me so strongly that I could focus on my research work.

At last thanks to the almighty God, who has blessed me with the spiritual support in the form of my Uncle (Mr. Satpal) and Aunt (Mrs. Kavita) who strengthened me to bring out the best version of me in all the odds that I came across during this journey, without them it would not have been possible.

Date: **(Chhavi Dhiman)**
Place: Delhi **2K16/Ph.D./EC/07**
Department of Electronics and Communication Engineering
Delhi Technological University,
Delhi-110042, India

ABSTRACT

The concept of an intelligent identification of human actions in videos is evolving as an active research area of computer vision and has covered a wide range of applications such as Ambient Assistive Living (AAL) [1], healthcare of elderly people [2], Intelligent Video surveillance systems [3], human-computer interfaces (HCI) [4] [5] , sports [6], event analysis, robotics [7], intrusion detection system [8], content-based video analysis [9], multimedia semantic annotation and indexing [10] etc. With the advent of technology and proliferating demand of society, automatic video sequence analysis based systems have become the need of the hour and their application in real life is helping to raise the standards of safety and security in society.

The performance of the intelligent human action identification system greatly depends on the type of input fed to the systems, and features extracted from the input data. Feature designing plays an important role in understanding the actions in videos. However, various environmental conditions such as lighting conditions, cluttered background, partial or complete occlusion, crowded scenes, different viewpoint of the camera, size, shape, appearance and complexity of human actions, badly affect the process of discriminating feature. Such challenges have always pushed forth researchers to explore new dimensions of the solution from vision-based to sensors, from 2D data to 3D data based Surveillance Systems, integrating multiple features, over the years. Various algorithms [11] [12] [13] [14] have been developed by the researchers, keeping different challenging scenarios in mind. Various real-time depth and skeleton based fall detection systems [15] [16] [17] [18] are developed considering the affordable range of the common user and practical challenges involved in video

analysis. In addition to this, deep architectures [19] have also foot-stepped in computer vision field and used for automatic assessment of Parkinson’s disease, AAL applications and many more. Therefore, this thesis investigates both two-dimensional: RGB and three-dimensional: Depth and Skeleton based human action identification methods using both traditional handcrafted features as well as deep features. The human action identification objective is mainly divided into three steps:

- The first step deals with *human silhouette extraction.* For different types of inputs different human silhouette extraction methods are used, which are listed as follows:
    i. For RGB video sequences, entropy based texture segmentation helps to segment the human silhouettes from background.
    ii. While dealing with depth images, human silhouette extraction process is accomplished by using global thresholding.
    iii. For skeleton data, joining of skeleton 3D coordinates generate the human poses for each frame.
- The second step is *feature extraction and representation* using both traditional and deep learning models. For different types of combination of inputs, features are extracted with four different approaches, given as follows.
    i. For RGB video sequences, feature vector is generated by combining global Spatial Distribution Gradient (SDGs) representation and Difference of Gaussian (DoG) based STIPs which are scale, rotation and translation invariant.

ii. For Depth and skeleton data, a robust feature vector is computed by using $\mathcal{R}$-transform and Zernike moments based human pose description, which is robust in terms of translation, rotation and scale variations.

iii. For RGB and Depth data, motion dynamics of an action is represented as Dynamic Images (DIs) based CNN features and geometrical view-invariant details of human poses are defined as deep HPM based features followed by learning of temporal information using LSTM model.

iv. For skeleton data, part-wise spatio-temporal CNN – RIAC Network-based 3D human action features are defined.

- The third step is *classification of human actions.* K-NN, SVM, and HMM are used to classify traditional handcrafted features. Weighted, max, average and multiply late fusion strategies are used for deep learning models.

The performance of each proposed action identification model is tested with various publicly available datasets and compared with earlier state-of-the-art algorithms. In addition to this, a novel Abnormal Human Action (AbHA) dataset is generated, while developing an automatic abnormal human action identification framework targeting the elderly health care and made publically available.

Finally, the research work is concluded followed by future research direction as well as possible future applications which are highlighted and discussed in detail.

LIST OF PUBLICATIONS

- **C. Dhiman,** 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.
- **C. Dhiman**, D. K. Vishwakarma "A review of state-of-the-art techniques of Abnormal Human Activity Recognition", **Engineering Applications of Artificial Intelligence**, Vol. 77, pp. 21-45. 2019.
- D. K. Vishwakarma, **C. Dhiman**, "A unified model for human activity recognition using spatial distribution of gradients and difference of Gaussian kernel" ***The Visual Computers***, Vol. 35, no. 11, pp. 1595–1613, 2019.
- **C. Dhiman**, D.K. Vishwakarma, "View-invariant Deep Architecture for Human Action Recognition using late fusion", *arXiv preprint [online] https://arxiv.org/abs/1912.03632v1*, 2019.
- **C. Dhiman,** D. K. Vishwakarma, P. Aggarwal, "Skeleton based Activity Recognition by Fusing Part-wise Spatio-temporal and Attention Driven Residues". *arXiv preprint [online] https://arxiv.org/abs/1912.00576v1*, 2019.
- **C. Dhiman** and D.K. Vishwakarma, "High dimensional Abnormal Human Activity Recognition Using Histogram Oriented Gradients and Zernike moments" in *IEEE International Conference on Computational Intelligence and Computing Research* (IEEE ICCIC17), Karumathampatti, India, December 14-16, 2017.
- **C. Dhiman** and D.K. Vishwakarma, "A Hybrid Multimodal Tracking System for boarder surveillance" in *IEEE International Conference of Soft Computing and Network Security* (ICSNS), Saravanampatti, TN, India, February 14-16, 2018.
- **C. Dhiman,** M. Saxena, and D. K. Vishwakarma, "Skeleton-based View Invariant Deep Features for Human Activity Recognition", in *IEEE Fifth*

International Conference on Multimedia Big Data (BigMM'19), Singapore, September 11-13, 2019.

- **C. Dhiman,** S. Gupta and D. K. Vishwakarma, "Deep Facial Expression Recognition using Transfer", in *IEEE International conference on Signal Processing, VLSI and Communication Engineering* (ICSPVCE), Delhi, India, March 28-30, 2019.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

With the advent of camera technology, nowadays, we are flooded with profuse amount of image and video data. Smart cameras are used everywhere for surveillance at public place, for recording recreational activities in education, homes, banks, shops, medical domain, in air and under water. With the proliferating number of recorded videos and their widespread availability, analysis of the video content becomes crucial in terms of social safety and security. Efficiency of human cognizance system starts getting effected while analysing videos for long durations. Therefore, automated systems must be developed to analyse and understand videos content efficiently for long hours. Identification of human actions in videos is an important area of video understanding. This thesis focuses on both handcrafted features based approaches and deep learning based feature extraction approaches to recognize human actions in videos.

This chapter introduces the fundamental building blocks of human action identification in video sequences, its application and the challenges involved therein. In the last section, the major research contributions of the thesis are discussed, followed by thesis organisation.

1.1 Human Action Identification

Human Action Recognition (HAR) broadly refers to identification of types of human action, activity, pose, gesture, automatically through computer or machine. Therefore, identification and recognition terms are used interchangeably in the thesis.

Identification of Human Actions in video sequences has appealed tremendous attention in the field of computer vision. It is also because of its gigantic potential in wide spectrum of areas i.e. Ambient Assistive Living (AAL) [20] [21] [22], healthcare of elderly people [23], Intelligent Video surveillance systems [24] [3], human-computer interfaces (HCI) [25] [26] [27], sports [28] [29], event analysis, robotics [7], intrusion detection system [30], content-based video analysis [31] [9], multimedia semantic annotation and indexing [10] etc. However, the real time videos project complex actions sometimes with large inter class similarity or large intra class variations, illumination variations, view variations, partial or complete occlusion, cluttered background and moving camera.

Before understanding the details of identification of types of Human Actions/Activity, let us first understand about human actions/activity. Human activity can be broadly categorised into four categories [32]: *gestures, actions, interactions and group activity.* Gestures are the fundamental motion patterns of the human body parts i.e. raising a leg, stretching an arm. Actions are defined as single person activities having a collection of multiple gestures i.e. walking, punching and waving. Interactions include at least two people and/or object such as two people fighting, a person stealing the suitcase. Group activities are performed by a group of people and/or objects i.e. group fighting, or a meeting. This thesis mainly focuses on human actions and its identification in video sequences. A vision based human action identification system automatically extracts the spatial and temporal information about the action from action sequences to recognize the action. The fundamental steps of human action identification in videos include pre-processing, feature extraction and classification, as shown in Fig. 1.1. In order to understand the real world complex actions and activities automatically, an identification system must have the understanding about the basic human poses like, standing, bending, walking, sitting down etc. Pre-processing of action sequences helps to detect the person in action in the video, and feature extraction stage ensures that meaningful spatio-temporal information of human poses, hence the action performed, is extracted. In classification stage, the identification system makes a final decision about the action on the basis of extracted features. Therefore, each stage of the identification system greatly effects the action recognition performance.

Pre-processing of RGB frames involve segmentation of human silhouettes which can be performed by using background differencing, background modelling, texture based segmentation, and optical flow based object segmentation etc. With the innovations in the field of imagery technology, over a period of time, it has become possible to capture the depth and skeleton representation of the actions in real time along with RGB representation in cost-effective manner. Due to this, the inputs to the Human Action Recognition (HAR) systems are now available in three forms namely RGB, Depth and Skeleton. It has expanded the dimensions of feature extraction approaches to represent an action. Since, depth value estimation involves infrared radiation, which are not sensitive to lighting conditions or visible light. They make foreground extraction much simpler and faster at pre-processing stage. Whereas skeletons provide access to 3D coordinates of human joints in 3D space. Due to these benefits of depth and skeleton images, researchers are now inclining towards depth and skeleton image based action analysis. Spatial representation of skeleton joints helps to observe the human pose variations of the action. The change in skeleton joints' location in 3D space over time provides temporal representation of the action. Researchers may either process all the frames of the action sequence or select only the distinct key poses from the action sequences, in order to reduce redundant data. With the availability of enormous amount of data, feature extraction methods are evolving from traditional handcrafted features to deep learning based features.

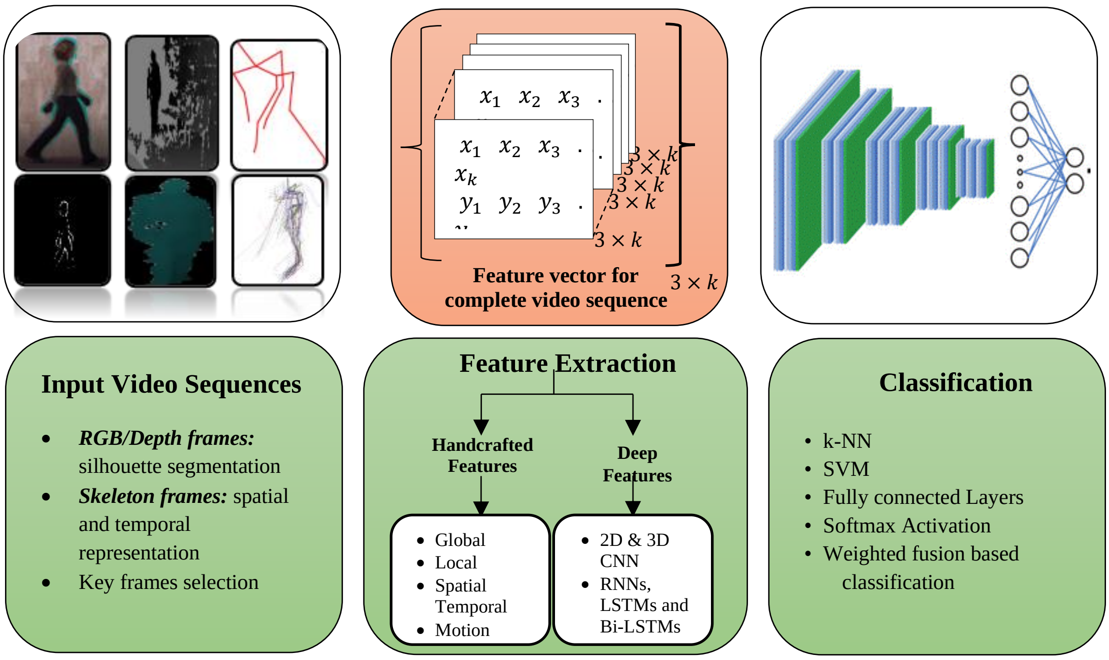


Figure 1.1: Overview of Human Action Identification (HAI) system

Deep learning approaches generate more robust and dynamic features than the traditional handcrafted features. It is due to the fact that deep learning approaches start from pixel level at layer one and look for all possible combinations of edges and shapes formed, layer by layer in the training phase. 2D Convolution Neural Networks (CNNs), 3D CNNs, two stream and three stream networks are some of the well-established approaches for deep spatial feature extraction. Whereas, Recurrent Neural Networks (RNNs), Bi-Directional Long Short Term Memory (Bi-LSTMs), Long Short Term Memory (LSTM), are used to learn the temporal changes in human poses. K-NN, SVM, HMM, Random Forest, Bootstrap are some of the popularly used traditional classification approaches. Whereas, Softmax is used for classification, in deep learning approaches. In multi-stream networks, early fusion and late fusion are preferred approaches for final prediction. Weighted fusion is one of the method to perform late fusion and early fusion deals with combining the multi-stream features which are later classified by applying Softmax layer on it.

1.2 Major Challenges

Despite the consistent efforts made in the domain of human action identification in video sequences, it is a challenging task to develop a discriminative action representation in realistic videos. A variety of environmental challenges exist such as presence of cluttered background, viewpoint variation, illumination variation, scale variations, high inter class similarity, high intra class dissimilarity, different colour and texture of clothes. The effective action identification system requires a good descriptor capable of handling these challenges.

The environmental conditions during video sequence recording greatly affect the performance of vision based recognition system. Different amount of illumination exposure or night vision distort the information acquired in images leading to the poor quality of video frames. Due to which, section of pixels in an image are highly illuminated and some of them appear dark and the information about the object in these sections gets unclear. Hence, the object and background of the scene may be indistinguishable. It affects the results obtained at pre-processing and feature extraction stages of action identification process. Therefore, in pre-processing stage, the effect of illumination must be reduced so that the spatial information of the object in the frame can be correctly detected. For this purpose, histogram equalisation [11] is used to normalise the contrast of the pixels but despite the added complexity to the system, histogram equalisation based approaches could not result in stable performance for different amount of illumination conditions. A few researchers [33] [34] processed texture data of the object to segment the object from the frame using GMM, LBPs etc. However, GMM and LBP based object segmentation involves significant amount of processing time. Nowadays, Microsoft Kinect camera provide depth information of the scene along with RGB information. Since, depth images are generated by IR rays which are insensitive to illuminations. Therefore, depth images are also illumination invariant. Hence, depth image based action recognition can be preferred solution.

The camera movement while acquiring the action, causes blurriness in the image that makes the boundary of objects in the scene hazy and unclear. Therefore, additional de-blurring algorithm needs to be applied. Hence, an action must be captured in a controlled environment with proper camera calibration.

The cluttered background increases the complexity of the background and makes the foreground extraction difficult which is responsible for inaccurate segmentation of the object. This is noted especially in outdoor activities, where the cluttered background cannot be avoided. Hence, complex foreground modelling approaches like Gaussian Mixture Modelling (GMM) [33] [34], salient object detection algorithms [35] are used to extract the object.

There is a high possibility that while performing an action, the object may get occluded partially or temporarily, that result in occlusion of significant human pose information. There are different types of occlusions, categorised as temporary occlusion, permanent occlusion, self-occlusion, partial occlusion and complete occlusion. Permanent and complete occlusions are most dangerous type of occlusion. Because they result in loss of spatial information of the object permanently, which we cannot retrieve during the entire video sequence. Whereas, in case of temporary, self and partial occlusions, the spatial information of the object under action is not visible only for few frames. The effect of occlusion can be reduced by using multiple camera recordings which provide information at multiple views [36], [37] for recorded actions. However, use of multiple cameras increases the cost and complexity of system.

The action performed at different angles from the camera, by different performers, is also an important challenge which greatly affect the performance of an automatic identification system. Variation in viewpoints of action leads to occlusion of the object. Therefore, to extract the occluded information, multiple views are used. The methods [36] have been evolved which extract unique features of the actions irrespective of the views, such that change in view does not affect the performance of the identification system.

1.3 Problem Statement

The challenges involved in identification of human actions in videos, motivate to develop efficient algorithms to fulfil the purpose of developing robust recognition system. In order to achieve this, four problem statements are formulated here to handle the practical challenges involved mentioned above, which are as follows:

- ✓ Design and development of translation, rotation and scale invariant framework for automatic identification of abnormal human actions in video sequences, with less execution time.

- ✓ Establish a novel framework for human action identification in RGB video sequences, which is robust against scale and illumination variations.

- ✓ To establish a novel view-invariant deep RGBD human action identification by integrating two important action cues: motion and shape temporal dynamics (STD).

- ✓ To design a novel part-wise spatio-temporal CNN – RIAC Network based 3D human action recognition framework.

1.4 Theoretical Formulation

- The need for an appropriate type of segmentation approach is identified, which can handle different illumination conditions, and spatio-temporal redundancy in the RGB video sequence.

- The issues involved in recognising human actions under different viewpoints are highlighted. A recognition system can understand view variations of the actions performed by different people properly, if it is introduced with possible multi-view human poses of the action in the learning phase.

- The challenges in identifying human actions correctly for a set of samples with large similarities among different action classes and small similarity among actions samples belonging to the same action class, are identified and addressed by using the concept of part-wise spatio-temporal CNN features.

- The use of depth images of action sequence simplifies foreground segmentation. The use of $\mathcal{R}$-transform and Zernike moments based human pose descriptor introduces translation, scale, and rotation invariant action description, which enhances the human action identification performance.

1.5 Experimental Validation

The developed algorithms are experimentally validated using publically available datasets. These datasets have numerous real life challenges such as illumination variation, cluttered background, scale variation, rotation and translation variations, high inter class similarity and intra class dissimilarity.

- Entropy based texture segmentation model exhibit illumination invariance for RGB videos sequences. The spatio-temporal redundancy is removed by converting the RGB video sequence into the compact Average Energy Silhouette Images (AESI) representation.

- Difference of Gaussian based STIPs and Spatial Distribution Gradient (DoG-SDG) action descriptor enriches the global shape description given by SDG with implicit scale and rotation invariant property of local STIPs.

- It is experimentally observed that unification of $\mathcal{R}$-Transform and Zernike moments preserves the translational, scale and rotational invariance of the action descriptor leading to improved performance with low computation cost.

- The Deep Part-wise Spatio-Temporal CNN – RIAC Network based 3D features are observed to be superior over the global deep RIAC-Network based features and validated for publically available standard datasets.

- Two stream deep network turns out to be robust against view-variations, which is enriched by two important action cues: motion and shape temporal dynamics (STD) for a RGBD action sequence.

1.6 Motivation for Human Action Identification

The key motivation behind the study of human action identification is its proliferating applications in real world including critical issues such as safety and security of humans in society. More than two-thirds of the accidents suffered by the elderly people are due to fall. It becomes rather difficult to leave the elderly people unattended alone at home. The European Statistical Office reported that by 2060, the ratio of young and old person will be 1:1 in the EU [38]. In addition to this, World Health Organization [39] mentioned that injuries due to fall will shoot up by 100% by 2030. Henceforth, the concept of smart homes, is appealing to the research community [40] [41] [42] [43]as a means to support peoples' health. Technological inventions and development play a significant role in extending support through assistive and autonomous care facilities, preferably based on a low-cost setup, that can be set up easily at homes and care centres.

A real time Virtual Exercise Rehabilitation Assistant-VERA is developed by a Reflexion Health organisation, San Diego, U.S. which using the concept of depth silhouettes based human poses analysis. It is a patient-centered, data-driven, and value-based solution for post-acute care rehabilitation. It is helping the patients and clinicians to efficiently manage physical therapy from pre-habilitation through to post-acute care. VERA, a digital platform, is currently used at academic medical centers, ambulatory surgical centers, home health organizations, senior living communities, and individual patient homes across the U.S.

Recently, a tele-robotic surgery is performed remotely by Dr. Tejas M. Patel using robotically controlled instruments at Ahmedabad-based Apex Heart Institute, Gujrat, India. The instructions given by the doctor with his gestures, and speech, were understood and executed by the robot successfully. It is one of the best examples of how Artificial Intelligence (AI) can contribute to the benefits of the mankind.

Such application of intelligent systems motivates us to develop real-time robust human action identification framework which can help to distinctly recognise, both normal and abnormal types of actions despite of translation, scale and view variations. From the existing literature [43] [44] [45] [46], it is identified that multiple features provide richer information than one type of features. Therefore, in this thesis, novel human action identification frameworks are established which unify more than one type of action features, to achieve higher recognition accuracy.

1.7 Significance of the Study

Human action identification has been a growing field in computer vision because of its gigantic potential in wide spectrum of areas i.e. Ambient Assistive Living (AAL), healthcare of elderly people, Intelligent Video surveillance systems, human computer interfaces (HCI), event analysis, sports, intrusion detection system, robotics content-based video analysis, home automation etc.

The key findings of this study leads to expand the dimensions of human action identification in video sequences to a wider framework for many real life applications. These applications will foster the security to a higher level in day-to-day routine of common user. Another important significance of this study is to establish a state-of-the-art that can escort the research community to dig deeper in this area.

1.8 Thesis Overview

Chapter 2 provides the detailed study of the existing state-of-the-arts and their analysis i.e. merits and demerits at different levels of video analysis: pre-processing, feature extraction and representation, and classification. It helped to draw an outline of research gaps in the concerned area. The final research objectives are defined which are addressed in the thesis later.

In Chapter 3, two handcrafted features based models are defined to recognise human actions in video sequences. The first model unifies spatial distribution of gradients and difference of Gaussian based STIPs, to represent the actions in RGB videos. The second model describes a robust framework for abnormal human action recognition using $\mathcal{R}$-Transform and Zernike moments in depth videos. The strengths of the proposed models are supported by detailed explanations of video pre-processing, feature extraction and representation, classifications, experimental setup and discussion of results.

Chapter 4 discusses about two deep learning models for human action identification in videos. The first deep model utilises the concept of transfer learning to develop view-invariant action representation by combining both RGB and depth frames of the video samples. The second proposed deep model highlights the strength of Part-wise Spatio Temporal attention driven CNN features for 3D human action identification in videos.

Lastly, chapter 5 highlights the important conclusions drawn from these methods and gives the details of future scope of work.

CHAPTER 2

LITERATURE REVIEW

This chapter discusses about the existing state-of-the-arts for robust human action identification. To understand the evolution of various types of action descriptors over time, which enhance the ability of recognition systems, the literature is broadly classified in two categories: Handcrafted features based approaches and Deep features based approaches. Both the categories cover the state-of-the-arts developed using 2-Dimensional RGB data and 3-Dimensional Depth and Skeleton data to recognise the actions to make the identification of actions more effective.

2.1 Traditional handcrafted features based approaches

Handcrafted features refer to manually designed features. The design of handcrafted features requires right selection of discriminant features and often involves a trade-off between accuracy and computational efficiency. A large number of features, as reported in the literatures [47] [48] [49] [12] have been manually designed either to extract texture, colour, shape, spatial information, scale and view invariant information, temporal information , motion information as spatio-temporal interest points (STIPs), Scale Invariant Feature Transform (SIFT) [50], Discrete Time Warping (DTW) [51], optical flow (OF) [52], Dense Trajectory based features (DTF) [12], respectively, to understand the close characteristics of actions. The selection of key feature also depends on the type of inputs available to the system which hold the action characteristics such as RGB video sequences, depth video sequences, skeleton representation of action. Choosing the right trade-off between accuracy and computational efficiency is crucial.

Human action identification (HAI) approaches are broadly categorized as single person HAI and Multiple Persons HAI. The thesis focuses on single person action recognition in videos, which are categorized into three sections on the basis of their significant contributions in the field of single person action recognition. First subsection, discusses about the approaches defining silhouette based and spatio-temporal features of the RGB action sequences. Second subsection deals with depth and skeleton based action representation methods.

2.1.1 RGB based HAI approches

A comprehensive survey [53] focused on the developed methods from 2008 to 2012. These human action identification approaches can be broadly divided into three categories: human detection (low-level vision), human tracking (intermediate-level vision), and behavior understanding methods (high-level vision). Another survey [53], analysed various scene behaviour modelling approaches which extracted context based features. Over the year's behavior modeling focus has shifted from rule-based methods to probability-based statistical methods, being superior in robustness and scalability. But the real challenge still lies in making the system free from human interventions and minimizing false alarm (positive/negative) rate. For which global and local features have proved to be highly descriptive and discriminative in nature to encode the behavior. In addition to it, trajectory, speed, and direction, optical flow, object-based abstraction methods are also used for complete abstraction of the scene behavior. Whereas, for high dimensional behavior modelling, learning with Generative Topic Model framework [53] generate more robust sparse spatial-temporal interest points than and can determine eloquent activities from co-occurrence of visual words automatically. Various vision-based approaches or RGB based human action identification approaches [54] [13] [55] [56] have utilised spatiotemporal details, shape deformation features, posture information to analyse the scenes. In this section various recognition methods in RGB videos are discussed under silhouette based and spatio-temporal information based approaches. Since AAL, home monitoring is one of the major application of single person HAI, third subsection highlights various HAI methodologies approaches supporting elderly health care issues and AAL [41].

2.1.1.1 Silhouette based approaches

Robustness of any HAI method depends on the extraction of action evidence from an image in an efficient manner. Human silhouette is a fundamental entity that provides a holistic description of the action. In many approaches [57] [58] [59], the foreground is extracted from the background to obtain human silhouette. In HAI, a set of features generated from image sequences, must be suitably rich for robust classification against illumination, occlusion, viewpoint, camera motion, compression and frame rates. In [60], the human silhouette is extracted in three steps: pre-processing, probability map as a weighted sum of mean shape template prior, colour histogram prior and saliency prior, and graph-cut segmentation. It proves to be an appropriate solution for still images, in terms of accuracy and computation time. A similar approach [61] combined shape and appearance based probabilities, computed for local windows inside the bounding box. Further, the Colour histogram probabilities followed by GMM probabilities are computed for each shape and appearance cue, which are then fused together by weight maps generated from the genetic algorithm. Segmentation cut is applied on the fused image to extract the silhouette. Since this approach depends on colour histogram probabilities, the segmentation results are sensitive to illumination variations. For RGB datasets, with clutter background and more texture, strong 3D gradient are extracted from a background which is mostly not associated to the action and acts as a noise. Additional de-noising methodology needs to be followed, which is responsible for additional time complexity. Laplacian fitting scheme [62] is also used for automatic extraction of human silhouette. The raw object motion segmented images are taken as input and Laplacian matrices are computed and minimized to obtain foreground human silhouettes. For part based human attribute recognition, heat maps [63] are preferred to localise the salient objects. Human segmentation process is a pre-processing step of video processing; therefore, it needs to be computationally less complex with decent segmentation results. It is observed that human segmentation approaches have improved the accuracy of background modelling and foreground segmentation in a reliable manner. However, due to the inherent complexity of real videos, human silhouettes extraction tends to suffer with high computation cost. Instead, the extraction of human silhouettes in many well-known human pose datasets [64] still rely on manual intervention with significant time consumption. In the proposed approach, the employed entropy based human silhouette segmentation for RGB video considers, both accuracy and computation cost parameters. Hence, fine human silhouette segmentation is performed with needful computation time, against the posed challenges of the datasets used.

R-Transform has been used frequently [65] [66] to target elderly health care issues and providing highly practical solutions. Khan and Sohn [66] considered six possible unusual activities (i) faint (ii) backward fall (iii) forward fall (iv) vomit (v) chest pain and (vi) a headache. The work integrated R-Transform with Kernel

Discriminant Analysis (KDA) to minimize interclass similarity of different activity postures as binary silhouette maintaining individual's privacy. Khan et al. [65] designed a two-level hierarchical anomalous human activity recognition system to increase the recognition rate for intra-class activities, particularly for falling forward - vomiting postures and falling backward - fainting postures. This hierarchical approach has increased the average recognition rate of similar activities to 97.1% in confined environment which needs to be extended with real-time abnormal activity dataset augmented with noise. Binary silhouette fails to describe the posture in case of self-occlusion. This limitation can be removed using depth silhouettes.

Shape analysis of human silhouettes plays a vital role to understand the action performed. Hence, a large number of researchers [54] [67] [68] have focused on shape analysis of a person. Rougier et al. [54] analysed the human shape deformations using shape context matching [67] to detect fall using Gaussian mixture model which can be used to confirm the safety of elderly individuals at home. Recently, recurrent neural network (RNN) based fall prediction method [68] analysed human biomechanics equilibrium with 91.7% fall prediction. It computed unbalanced posture features using skeleton joints. Yuna et al. [69] utilised R features to capture the geometrical statistics of the interest points, which are invariant to geometry transformation and robust to noise.

2.1.1.2 Spatio-temporal features based approaches

For human action identification, a remarkable amount of work [70] [71] [72] [73] has been reported by designing spatio-temporal features using both global and local evidences. The global evidences provide a holistic representation of the action in terms of shape and motion, whereas the local evidences highlight local details of the action and are less sensitive to noise. Motion Energy Image (MEI) and Motion History Image (MHI) introduced by Bobick and Davis [74] greatly influenced the holistic representation of the action. Recently, Ijjina et al. [75] modelled the motion information by defining temporal templates (TT) as a weighted sum of MHI and MEI of the RGB-D video. And this temporal representation of the video is given as input to CNN to predict the class labels. The temporal templates for each action is obtained from RGB and binarized depth stream. Binarized depth silhouettes are similar to binary silhouettes. But binarization leads to loss of depth details which may provide the fine variations of depth silhouette. Recently, human gait motion in a video is represented by 2D Spatio-temporal template [76], called as Average Energy Silhouette Image (AESI) that preserves all the shape variations with time, in a single frame with an additional advantage of storage and less time consumption in post-processing steps. It motivated us to employ SDG descriptor on AESI representation of an action video instead of each frame that makes the SDG descriptor computation for entire dataset faster and complimentary for real-time applications. In [77] shape and motion orientation of the object are used as baseline features to define action. The approach utilised spatial edge distribution of gradients and texture based segmentation technique to extract binary silhouette. Whereas, motion orientation is obtained with the help of R -transform followed by Local Linear Embedding (LLE). Gaglio et al. [78] defined human activity as spatiotemporal evolutions of different body postures. The recurrent postures are represented by most substantial structures of joint locations. Aggarwal et al. [76] applied Zernike moment invariants (ZMI), a shape descriptor, on Average Energy Silhouette Images (AESI) to detect undesired covariates in the gait sequences such as clothing or carrying bag. However, this approach needs to be tested for multi-view gait samples. Sintorn et al. [79] defined Regional Zernike Moments (RZM) that combine zernike moments of pixels confined in a region to create a suitable measure for texture analysis. In [80], content-based image retrieval complex zernike moments are used for shape feature extraction. A concept of discriminative and informative semantic for human action recognition is used in [81] to overcome the problem of non-extraction of abundant visual spatial-temporal information using local and global features, using mid-level representation based on optical flow method, Hu, and zernike moment together. Local representation of action emerged from the pioneering work of Laptev [82] on Spatio-temporal Interest key points (STIPs) detection which is later extended to local descriptor extraction and aggregation of local descriptors. Raptis and Soatto [83] introduced spatio-temporal action descriptors using HOG or HOF along the trajectories to represent the local appearance and geometrical information around trajectories. It modelled the actions by using the bag-of-words model. The descriptor is based on low-level statistics which do not enforce global shape, motion statistics. Hence, the obtained action recognition accuracy is not optimal for practical scenario. P. Lishen et al. [84] represented each action video as the histogram of visual words obtained by pooling learned local spatial features from action cubes. The performance of spatial features greatly depends on the pooling strategies. Hence, weighted pooling strategy should be followed. In [85], perpendicular local binary pattern (PLBP) is presented to describe textures in the local neighbourhood of a pixel efficiently by considering the relative differences of intensity between a pixel and its neighbours. It handles noise adaptive thresholding based on the image contrast of a region. However, performance suffers with the trade-off between accuracy and computational cost

2.1.1.3 Identification of actions for Elderly health care and AAL

A Few surveys [42] [86] are identified, discussing about computer vision solutions for elderly health care, home surveillance and AAL. Rashidi et al. [42] talked about Ambient-Assisted Living tools and techniques for the elderly people using various types of wearable and ambient sensors to vision sensors. Chaaraoui et al. [86] highlighted the challenges of sensor technologies, limited assistive robot technologies, social security and privacy issues of AAL systems to make it widely acceptable among users. Whereas, in [41] the focus is brought to IoT, wearable devices, cloud computing, advanced robotics, sensor networks based assistive living products to discern the wider frontiers of AAL for healthcare, rehabilitation and assistive living. A broad survey [53] of video-based anomaly detection has brought forward diversified work much more in-depth. It defined the characteristics of an anomaly and context-based anomaly which may not be an anomaly in another frame of reference. Various scene behavior modeling methods are defined, considering behavior abstraction. LOTAR framework [87] offers a stronger feature representation platform for AAL application which analyzes both short term and long term anomalies by collecting data from multiple sensors i.e. temperature, pressure & RFID sensors along with vision sensors. For experimentation, the framework is employed in real patient home, which needs to be extended to multiple individuals for realistic results. In work [88], the habit of the person is studied and analyzed for the first time by fusing ISUS (Intelligent space for understanding and service) and multi-camera positioning algorithm.

2.1.2 Depth and Skeleton based HAI approaches

For 3-Dimensional HAI, the system is fed with depth silhouettes and skeleton structures of a person. Both kinds of approaches have their own set of applications and benefits. Recently in [89], an overview of various available depth sensors and their benefits over conventional cameras are described. It is observed that the growing research area is addressing human action recognition as normal or abnormal and focusing on depth based body part detection, pose estimation, body pose modeling, and space-time evidences. Some surveys [89] [90] [91] [92] conferred about depth imagery based human motion analysis, 3D skeleton based human action classification and introduced new datasets for handling complex interactions and smart home activities respectively. Depth images not only simplify and fasten up the low-level image processing but also deliver better processing outcomes in terms of background subtraction, object motion detection, and localization. In the following series various popular depth based approaches are discussed.

2.1.2.1 Depth based action description

Jalal et al. [93] defined the human pose feature as HOG-DDS which represent the projections of the Depth Differential Silhouettes (DDS) between two consecutive frames onto three orthogonal planes by the histogram of oriented gradients (HOG) format. Further, it is fused with skeletal based key joint-based distance feature (DK), the spatiotemporal magnitude feature (M), and the spatiotemporal directional angle feature (θ) and torso based distance feature (DT). For IM-DailyDepthActivity dataset, the body shape feature HOG-DDS and skeletal features-$\{DK, DT, M\}$ individual1y obtained 45.12% and 51.70% recognition accuracy, as reported in [22], which indicates the considered features are not efficient descriptor of the action. In [94], a supervised spatio-temporal kernel descriptor is defined to represent the RGB-D action.

Depth-silhouette based statistics such as height to width ratio, centroid, and silhouette shape deformations, are commonly used to extract features of the person under motion. Human motion and shape variation features [54] [13] can handle realistic challenges such as occlusion, different viewpoints etc. A new dataset CONVERSE representing Complex Conversational Interactions between two individuals via 3D poses in the survey has opened more possibilities for Abnormal Human activity Recognition (AbHAR). This dataset caters real-world challenging scenarios incorporating frequent primitive actions, interactions, and motion over a period of time. It is quite evident that in this decade, it is the vantage point for posed based 3D abnormal human activity recognition in research. Presti et al. [90] discussed different aspects of data pre-processing, publically available benchmarks, and commonly used accuracy measurements along with feature representation and 3D Skeleton based action classification at length. The concept of inactive period strengthens the severity of fall. A person who is lying on the floor and is inactive for an extended amount of time indicates a severe fall. Fall detection cannot be made from one instance information, but discriminative features need to be analyzed for the entire duration of fall and also after it. The confirmation of inactivity is highly context dependent. The exact location of person, time and duration of inactivity collectively leads to sensible decision such as staying in bed for long hours is not an alarming event but staying on the floor for long, after a fall will lead to an alarm. Hence contextual information helps to reduce the false alarming rate. Therefore, Jansen et al. [55] learnt about contextual details of the fall by quantifying the area of body on the floor and 3D orientation of fall to understand the inactive duration of a person during and after fall. In daily activities, human body undergoes various quick movements which may lead to false large motion identification. Therefore in [13] center of mass of human 2D silhouette is quantified to observe overall motion (magnitude and orientation) of the object, with the help of image moments, along with human shape feature. It reduces the impact of sudden movements on fall detection. In addition to this, MHI (motion history image) provides exact location and trajectory of motion in the video sequence. Yao et al. [95] introduced Human Torso Motion Model (HTMM) which can discriminate fall and fall-like activities such as bending and crouching down with 97.5% accuracy by observing changing rates of torso angle and the centroid height. Since the existing RGB-D action datasets i.e. CAD-60/120 do not provide fall sequences, ADL and fall sequences are recorded for experimentation. However, the method is dependent on threshold values obtained with trial and error approach to optimize the result, which needs to be identified every time for a new dataset. Rougier et al. [96] computed human centroid height relative to the ground and 3D person velocity. 3D person velocity helps the system to make fine discrimination between crouching down behind the sofa from fall behind the sofa – look alike cases. Here, 3D velocity is preferred over 2D velocity of a person during fall because 2D velocity is, generally, very high near the camera for normal walking activity resulting in misclassification between a fall and a walk. It is observed that height [97]and height velocity [98] based approaches fail to distinguish fall and fall-like actions and whereas bounding box width to height ratio based [13] [96] and HTMM [95] based fall detection model has the higher discrimintaion power in fall-like actions. Ma et al. [15] represented the actions by a bag of words model (BoCSS) using distinctive Curvature Scale Space (CSS) features of depth silhouette for fall detection, whereas Akagunduz et al. [16] integrated orientation scale space (OSS) and morphological scale space of a curve to form robust Silhouette Orientation volume (SOV) global scale invariant descriptor to represent actions. However, these approaches have high computational cost. Some researchers came up with fusion of sensor (accelerometer, floor sensor) and visual depth data which improved the performance of the fall detection. B. U. Toreyin et.al. [99] integrated sound impact of falling person with height to width ratio of the bounding box on a person under falling condition to discriminate a fall from a normal sitting in the floor action. Zerrouki et al. [100] used the concept of Univariate Statistical monitoring method Exponentially Weighted Moving Average (EWMA) control scheme to detect potential fall integrating accelerometric data and depth data with low computational cost. Though such fusions produce impressive outcome but this detection is dependent on sensor and its periphery which may not be in the comfort zone of the user. Therefore, the purpose of making the fall detection system non-intrusive to the user is defeated.

2.1.2.2 Skeleton based action description

Skeleton representation of human body provides incisive details about the human posture in compact form. This resolved the problem of need of effective segmentation technique to extract 2D silhouettes and simplifies the height centroid computation [96] from depth silhouettes. This, also, encouraged researchers to develop real-time applications using skeleton modality making the computation process faster, simpler and more effective. Nar et al. [101] designed an effective real-time ATM intelligent monitoring system to recognize abnormal postures prevailing stronger security in the ATM i.e. fiddling with the camera, aggressive posture, and peeping. The work used angles between different bones as useful features to compute the optimum weights' values to obtain the probability of current pose of the person under surveillance being abnormal. The computation of angle between joints (x, y, z) becomes quite simple, fast and more accurate with 3D skeleton coordinates. Hendryli et al. [102] addressed the issue of automatic detection of abnormal activities of students in examination hall that generates warning to exam proctors if any suspicious activity is detected (Cheating activity). For this purpose, MCMCLDA (Multi-class Markov Chain Latent Dirichlet Allocation) framework is designed that access arm joints and head location as interest points directly from skeleton representation without considering irrelevant ones resulting better accuracy and higher computational speed than Harris3D detector. Chaaraoui et al. [103] developed generic machine learning framework (Bag- of-Key-Poses) using joint motion history feature i.e. 3D location of skeletal joints and motion cues. To handle complex behaviors, both low and high-level multi-scale motion cues are extracted in [104]. However, skeleton data acquired with a Kinect sensor, is likely to suffer from a large amount of noise, and also contain outliers, especially in case of partial occlusion. Therefore, the work incorporates diffusion maps to filter the outliers. Jalal et al. [105], tried to develop a design continuous surveillance and daily activity recognition in indoor environments (i.e., smart homes, smart office and smart hospitals) turning the space into a smart living space. However, it failed to handle complex activities or partial occlusion of the body while generating skeleton joints and resulting in noise. To improve and expedite the medical care, an accurate automatic fall detector [106] is an essential element. Skeletal representation of a person is proving its strength by enhancing the performances of fall detection systems and many other abnormal activities in daily routine. Trajectory of joints [107] [108] [109] and Joint Motion History (JMH) [103] based action description is simple and effective with high temporal efficiency, appreciable view and illumination invariance property for skeleton-based abnormal human action detection. However, the distance between silhouette center and the floor [96] or shape deformation based fall detection work is not able to discriminate the initiative action from the fall accidents well i.e. fall in bed and fall on the floor without defining normal inactivity zones, a person is sleeping on the sofa or bed and falls down to the floor. 3D human skeleton joins distance from the floor, joins hitting velocity, joint position and its height from the ground [108] [109] collectively elicit robust results by discriminating a fall from slowly lying down on the floor and other similar cases. While falling, human body orientation changes dramatically which leads to poor tracking of joints. Therefore in [110], the author initially corrected the trunk orientation (from hip point to neck) of the person before applying a fast Randomized Decision Forest (RDF) algorithm for human skeleton extraction which has improved the accuracy of fall detection. The presented work is able to detect minor fall like falling from the sofa when our lower half body is still on the sofa by simultaneously tracking head, which silhouette center-based approach fails to identify. A view independent statistical method [17] takes a decision based on human behaviour information in last few frames after falling. It defined a feature set f = [f_1, $f_{2,}$ $f_{3,}$ $f_{4,}$ f_5], where f_1 is duration of fall, f_2 is total head drop, f_3 is the maximum speed of fall, f_4 is the smallest head height, f_5 is the fraction of frames where head has a smaller height than in the previous frame. The five features are combined using the Bayesian network. But few false alarms are received for a person lying on the floor and a person jumping on the bed. Diraco et al. [18] used distance of 3D centroid from floor plane as threshold to confirm high performance in terms of consistency and competence on a large real dataset which uses Bayesian segmentation to detect moving regions.

2.2 Deep features based approaches

It is observed that over the passage of time the concept of manual feature engineering is evolving from 2D features to 3D features in order to improve the action representation. However, complexity involved in designing handcrafted features is very high which is one of the key reasons to shift the feature designing methods from shallow region to deep, thereby boosting the practical applicability of the action recognition algorithms to higher level of excellence by empowering the knowledge of deep learning to recognition systems. Though the concept of deep learning and architectures [111] exist, since 1980s but they could not perform up to the mark due to the lack of sufficient datasets and computational resources. In1998, LeNet [112] came up as the first real-world successful application of CNN for handwritten digit recognition. However, in successive years, various deeper architectures have been reported in [113] and are being used in different application areas such as computer vision [114] [115] [116] , speech recognition [117], brain-computer interaction [118] and natural language processing [119] with the availability of large datasets and hardware resources. Deep models construct and learn from low-level features to high-level features. CNN is a type of deep model which is made up of neurons and learnable weights and biases which were initially applied on 2D images for visual object segmentation [120] and recognition [121] [122] tasks. And later, many researchers experimented CNN with videos by considering video frames as still images to recognise action in each frame. However, this approach was limited to learn only spatial information. Some authors [123] tried to incorporate temporal information by expanding the 2D CNN to 3D CNN. 3D CNN applies 3D convolution in CNN convolution layers by using 3D kernel to multiple contiguous frames stacked together to encode the motion information with spatial one.

In subsequent years, the work has extended, further by bringing the concept of multi-stream CNNs [124] based action recognition, which has fortified feature description of an action to a higher level. It makes the deep recognition system to analyse not just raw images at a time but also multiple set of inputs such as: RGB image, optical flow [124], dynamic images [125], and depth images [126]. Whereas, LSTM [127] has emerged as one of the most popular unsupervised model that learns temporal arrangements of frames and predict time series data.

It is observed from the previous works [128] [129] that appearance, motion and temporal information act as important cues to understand human actions in an effective manner [130]. The multi-stream architectures [131]: two streams [132] and three streams [133] have boosted the response of CNN based recognition systems, by jointly exploiting RGB and depth based appearance and motion content of actions. Optical flow [134] and dense trajectories [135] are majorly used to represent the motion of the object in videos. However, these techniques are not fine-tuned to include viewpoint invariance. Dense trajectories are sensitive to camera views and do not include explicit human pose details during the action. Depth human pose can be useful to understand the temporal structure and global motion of human gait for more accurate recognition. Recently, the skeleton-based action recognition approaches [136] [137] are progressing towards the temporal dynamics of the action using RNNs and LSTMs. Du et al. [136] encoded relative motion between skeleton joints which are split into anatomically relevant parts and passed through each independent subnet to extract local features. Shahroudy et al. [137] introduced a part-aware LSTM which possess part-based memory sub-cells and a new gating mechanism, showing the superior performance of LSTM over some hand-crafted features and RNN. To learn the human motion features of the skeleton sequence, RNN-LSTM [138] allows the network to access and store long-range contextual information of a skeleton sequence. Several authors [139] [140] exploited feature learning ability of CNNs which largely focused on a better skeletal representation and learning with simple CNNs. To better capture the Spatio-temporal dynamics of the skeleton sequences, some authors [141] [142] [143] used CNN as a spatial feature extractor and unified with RNN-LSTM network to model human motion. However, it is noticed that RNN-LSTM based approaches performed better. On the other side, the use of RNNs results in overfitting if the number of input features are short enough to train the network and computational time dynamically increases with the number of input vectors. Spatial-temporal encryption of skeleton action sequences is more descriptive than using just temporal information of skeletons-based action representation. Tu et al. [144] defined the correlation among three-dimensional signal using 3DCNN to capture spatial and temporal information of the action sequence. Liu et al. [145] mapped the skeleton joints in 3D coordinate space before extracting view-invariant Spatio-Temporal features, which significantly improved the action recognition results. Whereas, the work [146] learnt adequate geometric features of 3D human actions by using Lie Group and unified it with deep neural networks. Chen et al. [147] encoded the skeleton joints as part based 5D feature vector, to identify the most relevant joints of the skeleton during the action sequence using a two-level hierarchical framework. Amor et al. [148] used trajectories on Kendall's shape manifolds to model the dynamics of human skeleton poses and used a parametrization-invariant metric for aligning, comparing, and modelling skeleton joint trajectories, to deal with the noise caused by different execution rates of the actions performed. However, this method is time-consuming. A good amount of work is also done to address the spatial representation of human skeleton poses which are characterized by the interactions or combinations of a subset of skeleton joints [138]. The methods to model action spatial patterns can be categorised in two classes: part-based model and sub-pose model. In the first category of spatial pattern modelling, the skeleton is divided into several groups, instead of considering the complete skeleton. The HBRNN [149] model decomposed the skeletons into five parts, two arms, two legs, and one torso, and built a hierarchical recurrent neural network to model the relationship among these parts. Similarly, Shahroudy et al. [137] proposed a part-aware LSTM model that constructs the relationship between body parts. Whereas, in sub-pose model, the informative joints or their interactions are mainly focused. A handcraft feature based approach [150] defined a SMIJ model which selects the most informative joints by calculating statistical parameters such as mean and variance of joint angle trajectories and used the sequence of selected informative joints to represent the action. Wang et al. [151] mined co-occurrence distinctive spatial body-part structures as spatial part-sets and temporal evolutions of the pose as temporal part sets. Whereas Lillo et al. [152] learnt the Spatio-temporal annotations of complex actions using motion poselets and actionlet dictionaries. Such annotations help to understand which body part is active for a particular action but not discriminative enough in classification.

2.3 Research Gaps

On the basis of the outlines of literature review of the earlier state-of-the-arts for identifications of human actions in video sequences (RGB, Depth, or skeleton) research gaps are identified and a layout of solutions for the identified research gaps are listed, which are as follows:

- Identification of abnormal actions especially for elderly healthcare, demands accurate and automatic recognition. Therefore, a translation, rotation and scale invariant framework is designed for automatic identification of abnormal human actions in video sequences with low computational cost using depth videos.

- It is observed that scale and illumination variations badly affect the performance of action identification in videos. Therefore, a novel framework is designed which is robust against scale and illumination variations. The entropy based texture segmentation of human silhouettes introduces illumination invariance, and Difference of Gaussian (DoG) based Spatial Temporal Interest Points (STIPs) impart scale and view invariance to global Spatial Distribution Gradient (SGD) descriptor.

- Actions performed at different angles limit the performance of the action identification system greatly. Therefore, a view-invariant deep architecture is defined for human action identification using late fusion which learns the adequate multi-view human poses to correctly identify the actions irrespective of the view.

- The performance of an identification algorithm decays if the actions possess large inter-class similarity and large intra-class dissimilarity. Therefore, to handle these challenges, we have developed a Part-wise Spatio-temporal Attention Driven CNN based 3D Human Action Identification framework which reduces the inter class similarity and increases the intra class similarity resulting in improved human action identification performance.

2.4 Research Objectives

The main objective of the thesis is to analyse the practical challenges involved in the human action recognition in video sequences such as illumination, view variation, inter-class similarity and intra-class variations and further, to propose robust and computationally efficient action recognition frameworks. In order to achieve these objectives, the following frameworks have been proposed:

- A novel human abnormal action identification framework is defined which unifies the translation, rotation and scale invariant properties of $\mathscr{R}$ -transform and Zernike moments over structural appearance of human pose and its temporal motion content of an action, encrypted as Average Energy Silhouette Images (AESI). Human poses in depth videos are extracted as binary silhouettes by superimposing skeleton joints on depth images supporting low storage capacity and low computational cost.

- A hybrid framework for human action recognition in RGB video sequences is developed by integrating a set of global, local handcrafted features computed, which is robust against illumination variation, view variations via entropy based texture segmentation, and integration of view invariant Spatial Distribution Gradient (SGD) descriptor and Difference of Gaussian (DoG) based Spatial Temporal Interest Points respectively.

- A view-invariant deep human action recognition framework is proposed as a novel integration of two important action cues: motion and shape temporal dynamics (STD) by late fusion. The motion stream encapsulates the motion content of action as RGB Dynamic Images (RGB-DIs) which are processed by the fine-tuned InceptionV3 model. The STD stream learns long-term view-invariant shape dynamics of action using human pose model (HPM) [36] based view-invariant features mined from structural similarity index matrix (SSIM) based key depth human pose frames.

- A novel skeleton based part-wise spatio-temporal CNN – RIAC Network based 3D human action recognition framework is presented to visualise the action dynamics in part wise manner and utilise each part for action recognition by applying weighted late fusion mechanism. Part-wise skeleton based motion dynamics helps to highlight local features of the skeleton which is performed by partitioning the complete skeleton in five parts.

CHAPTER 3

HANDCRAFTED FEATURES BASED MODELS

This chapter introduces two handcrafted features based human action identification models using both RGB and depth videos in order to handle the practical challenges involved in video analysis such as scale, rotation, translation and illumination variations. The proposed frameworks are supported by the feature extraction and representation, experimental setting, comparative analysis of result, and discussions.

3.1 Abnormal Human Action Recognition Framework using $\mathcal{R}$-Transform and Zernike Moments in Depth Videos

This chapter presents a novel human action identification approach for elderly people. The most likely abnormal actions occurring with elderly people such as fainting, chest pain, headache, falling forward and backward, are analysed in order to reduce the dependency of the elders over others. The framework is structured to construct a robust feature vector by computing $\mathcal{R}$-transform and Zernike moments on Average Energy Silhouette Images (AESIs). The AESIs are generated by the integral sum of the segmented silhouettes obtained from the Microsoft's Kinect sensor v1. The proposed feature descriptor possesses scale, translation and rotation invariant properties, which is less sensitive to noise and minimizes data redundancy. It enhances proposed algorithm's robustness and makes the classification process more efficient. The proposed work is validated on a new abnormal human action (AbHA) dataset and three publically available 3D datasets - UR fall detection dataset, Kinect Activity Recognition Dataset (KARD) and multi-view NUCLA dataset. The proposed framework exhibits superior results from other state-of-the-art methods in terms of Average Recognition Accuracy (ARA).

3.1.1 Proposed Methodology

In the proposed work, a robust action descriptor is defined by combining the scale, translation and rotation properties of $\mathcal{R}$–Transform and Zernike moment. The action is described as a sequence of depth images which is later represented as AESI images encrypted by $\mathcal{R}$–Transform and Zernike moments. The proposed methodology is as shown in Fig. 3.1.

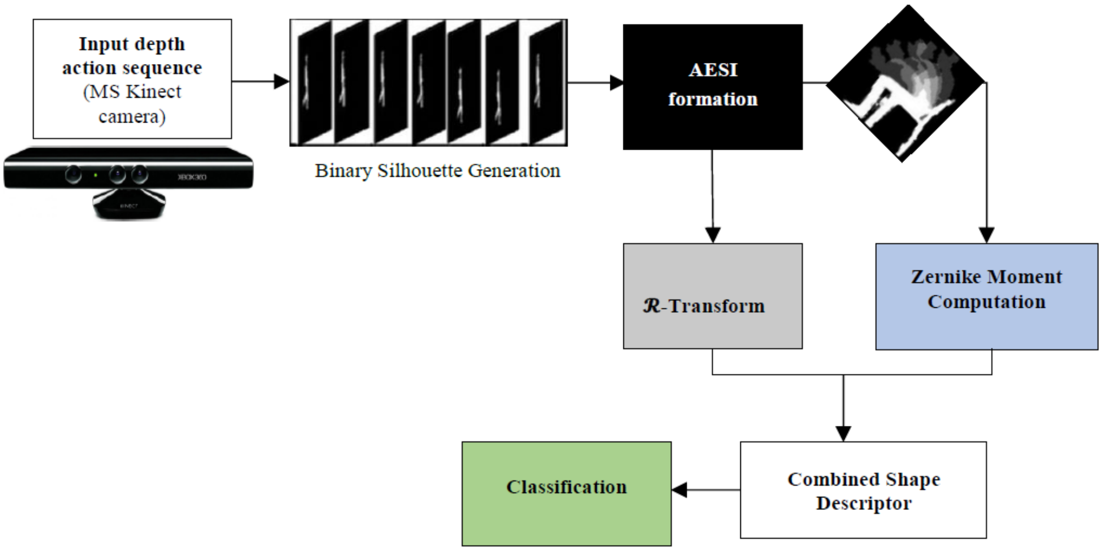

Input depth action sequence (MS Kinect camera)
Binary Silhouette Generation
AESI formation
R-Transform
Zernike Moment Computation
Combined Shape Descriptor
Classification

Figure 3.1: Flow diagram of proposed framework

3.1.1.1 Average Energy Silhouette Image (AESI) formation

Initially, fusion of depth action sequences acquired by Microsoft Kinect camera v1, and skeleton joint locations per frame, help to locate the person in the frame that makes the fine binary silhouette extraction easier. The entire action video is encrypted as a single image i.e. Average energy Silhouette Image, using extracted binary silhouettes. Consider any action with N key frames. Mathematically, AESI of an action is defined as shown in Eq. (3.1):

$$A(x,y) = \frac{\sum_{t=1}^{N} |I(x,y,t)|^{2}}{N} \tag{3.1}$$

where $I(x,y,t)$ represents each binary silhouette frame of the action at time instance$t$. $A(x,y)$ is the final AESI image constructed for one action sequence. The single frame representation of an action sequence as AESI not only removes computational complexity involved in processing entire video but the averaging function of AESI formation smoothens the unwanted noise in the frames. Hence, AESI images are less sensitive to the noise occurring between the frames, also termed as temporal noise.

3.1.1.2 $\mathcal{R}$ –Transform and Zernike moments based Shape Descriptor

Shape of the object possesses meaningful and important information about the action. A strong shape description is a key to a stronger recognition system. AESI is used as a compact representation of the action that holds the shape and its variations with time. Combination of $\mathcal{R}$ -transform and Zernike moments is used to define the shape descriptor.

***$\mathcal{R}$* -Transform :** $\mathcal{R}$ -transform [77] provides orientation detail of an object which is calculated by applying Radon transform, $\mathbb{R}_T$ on AESI images,$A(x,y)$. Radon transform generate directional features as the integral sum of $A(x,y)$, mathematically defined as Eq. (3.2):

$$\mathbb{R}_T(\sigma,\theta) = \iint_{-\infty}^{\infty} A(x,y)\delta(\sigma - xcos\theta - ysin\theta)dxdy \quad (3.2)$$

$$\sigma = xcos\theta + ysin\theta,\ \sigma\epsilon[-\infty,\infty] \quad (3.3)$$

where $\theta\epsilon[0,\pi]$, $\delta(.)$is a standard direct delta function which remains zero except at origin and σ is the shortest distance between origin and radon line, given by Eq. (3.3), graphically illustrated in Fig. 3.2. Radon projection, $\mathbb{R}_T$of an image cannot preserve all

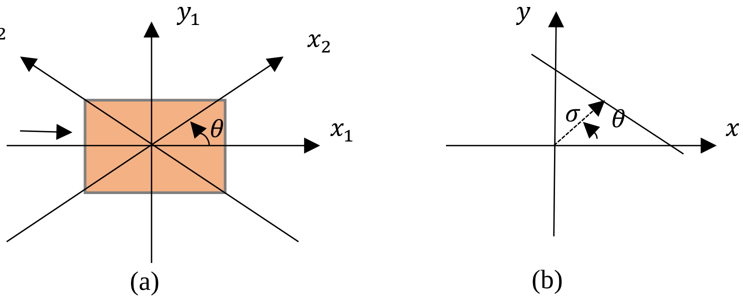


Figure 3.2 (a) Projection of radon lines over a 2D Image $A(x,y)$ (b) radon line

the characteristics of original geometric transformation- translation, rotation, or scaling the image. Tabbone et al. [153] introduced $\mathcal{R}$-transform that is defined as an integral transform of the squared values of Radon transform, $\mathbb{R}_T$, mathematically defined as in Eq. (3.4).

$$\mathcal{R}\,(\theta) = \int_{-\infty}^{\infty} \mathbb{R}_T(\sigma,\theta)^2\,\partial\sigma \quad (3.4)$$

where σ is the radial distance from the centre of the image to the radon line, and θ is the angle. Therefore, Radon transform $\mathbb{R}_T$ generates a 2-D feature representation and $\mathcal{R}$ -transform is a 1-D compact representation of $\mathbb{R}_T$. Normalization of $\mathcal{R}$-transform further improve the similarity measure of the feature vector, mathematically given as Eq. (3.5).

$$\mathcal{R}_{norm}(\theta) = \frac{\int_{-\infty}^{\infty} \mathcal{R}(\theta) d\theta}{max\,(\mathcal{R}(\theta))} \tag{3.5}$$

ℛ-Transform Properties**:** Tabbone et al. [153] illustrated the basic properties of $\mathcal{R}$ -Transform which confirms that it is scaling and translation invariant but sensitive to rotational characteristics. The properties of $\mathcal{R}$ -Transform are verified by considering chest pain activity AESI from AbHA dataset as shown in Fig. 3.3. It is observed from Fig. 3.4, case 4, that rotation in the original image leads to more changes in the pixel values falling on the projection lines, resulting in deformation in $\mathcal{R}$ -transformed image representation. The $\mathcal{R}$ -transform representation of different abnormal actions – ‘chest pain’, ‘headache’, ‘fainting’, ‘falling forward’ and ‘falling backward’, is shown in Fig. 3.4.

Zernike Moment: Ordinary geometric moments contain a lot of data redundancy. This is not desirable for any feature vector. Whereas Zernike moments reduces data redundancy with the help of complex polynomials called Zernike polynomials which comprise of a complete orthogonal basis set defined on a unit circle. This orthogonality of the polynomials helps to reduce the redundancy, hence producing an enhanced feature. Zernike polynomials [154] originally used to describe optical irregularities. For Zernike moment’s computation, we need to transform the image from Cartesian to polar coordinates. Complex Zernike moments can be determined by Eq. (3.6).

$$Z_{n,m} = \frac{n+1}{\pi}\sum_{(\sigma,\theta)\varepsilon unit\ disc}\sum f(\sigma,\theta)V^{*}_{n,m}(\sigma,\theta) \tag{3.6}$$

where, $V_{nm}(\sigma,\theta)$ is the basis function which can be determined by Eq. (3.7) followed by Eq. (3.8).

$$V_{nm}(\sigma,\theta) = R_{nm}(\sigma)e^{jm\theta} \tag{3.7}$$

$$R_{nm}(\sigma) = \sum_{s=0}^{\frac{n-|m|}{2}} \frac{-1^{s}(n-s)!\sigma^{n-2s}}{s!(\frac{n+|m|}{2}-s)!(\frac{n-|m|}{2}-s)!} \tag{3.8}$$

where, n is the order of the moment and m is the repetition. There are some constraints on the values of m and n: $n \geq 0, n \geq |m|,\ n-|m|$ is an even number. Since, the moments are calculated on the unit circle, $x^2 + y^2 \leq 1$ holds true. The incentive of using Zernike moment is that it is rotation invariant. Table 3.1 illustrates rotation invariant property of Zernike moment, which is unexploited in $\mathcal{R}$-transform. Lower order Zernike moments suggest the static information about the image by giving the basic details. Whereas, higher order moments describe the dynamic information by giving the image detailing. Jin et al. [155] established Zernike moments shape description efficacy by detecting heavily occluded targets such as airplane and warship efficiently.

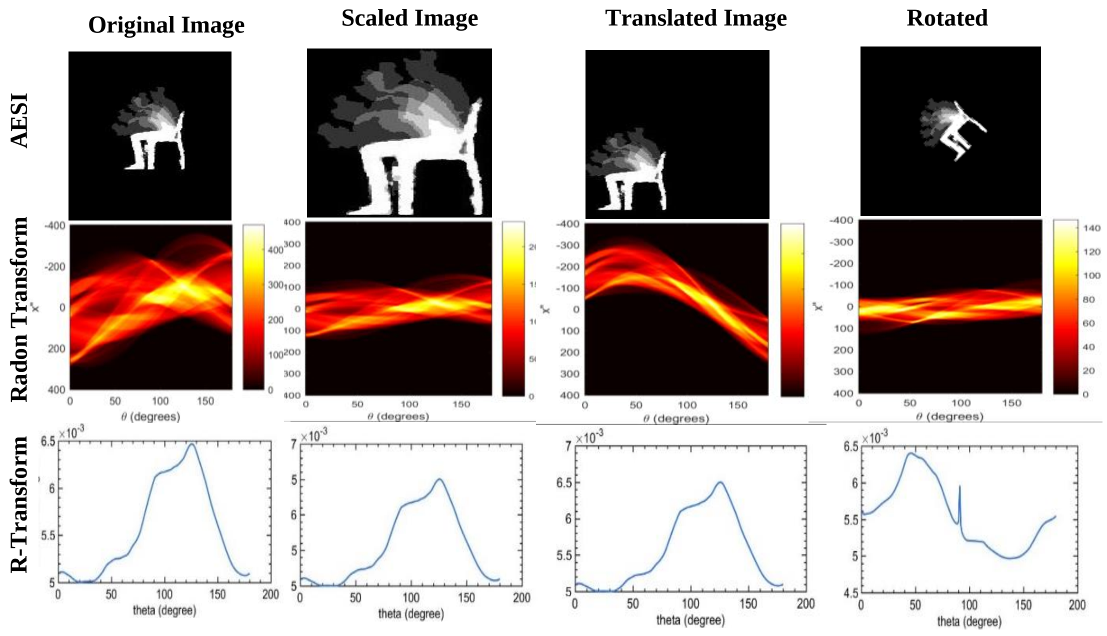
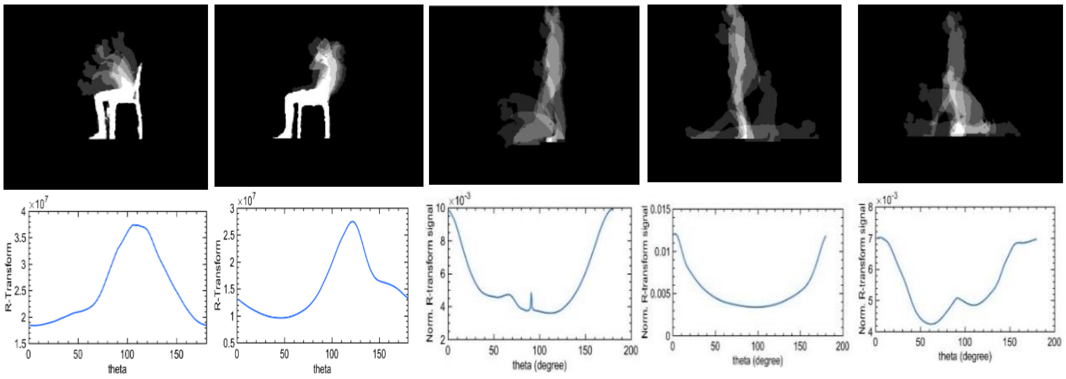


Figure 3.3: Illustrates that $\mathcal{R}$-Transform is scale and translation invariant but rotation variant



(a) Chest (b) Headache (c) Faintin (d) Falling (e) Falling

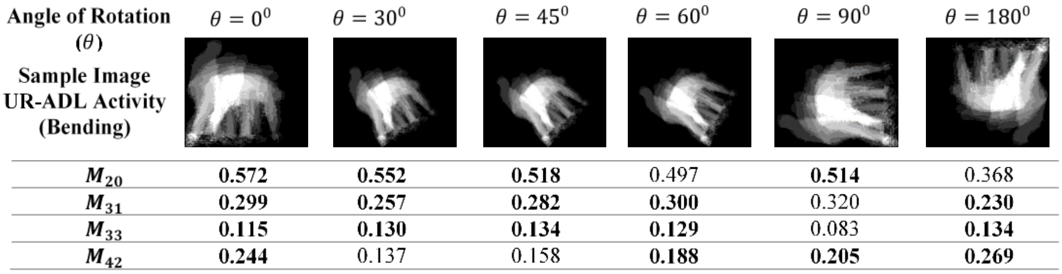
Figure 3.4: Representation of $\mathcal{R}$ Transform for five activities- AbHA dataset (a) chest pain (b) Headache (c) fainting (d) falling backward (e) falling forward. Row 1: 130×100 AESI, Row 2: Normalized $\mathcal{R}$-Transform

3.1.1.3 Final Feature Vector Formation

Radon Transform computes a 2D Projection of AESI of size $[480 \times 640]$ along the angle $\in (0^0, 179^0)$. It generates $\mathbb{R}_\mathrm{T}$ feature matrix of size$[803 \times 180]$. Application of $\mathcal{R}$-Transform develops a feature vector of size $[1 \times 180]$ independent of σ by integral sum of the squared values of $\mathbb{R}_\mathrm{T}$, Radon transform, as shown in

Equation (3.4) and simultaneously perform dimensionality reduction. Hence, $\mathcal{R}$-Transform produces 1–D projection of $\mathbb{R}_{\mathrm{T}}$ feature matrix given by $\mathrm{F}_{\mathcal{R}}$. Zernike moment feature vector F_{zm}, is obtained that includes one magnitude and phase value for every AESI image. Therefore, size of F_{zm} vector is $[1 \times 2]$. The final feature vector is defined by integrating $\mathbb{R}$ - transformed feature vector $F_{\mathbb{R}P}$ and Zernike feature vector F_{zm}. Final feature so formed possesses the dimensions as $[1 \times 182]$ per action. By doing so, we are able to inherit the desirable properties of both $F_{\mathbb{R}P}$ and F_{zm}

Table 3.1: Illustrates that magnitude of each Zernike moment is invariant under rotation. M_{nm}: magnitude of Zernike moment for order m and n.



Angle of Rotation (θ)	$\theta = 0^0$	$\theta = 30^0$	$\theta = 45^0$	$\theta = 60^0$	$\theta = 90^0$	$\theta = 180^0$
Sample Image UR-ADL Activity (Bending)						
M_{20}	**0.572**	**0.552**	**0.518**	0.497	**0.514**	0.368
M_{31}	**0.299**	**0.257**	**0.282**	**0.300**	0.320	**0.230**
M_{33}	**0.115**	**0.130**	**0.134**	**0.129**	0.083	**0.134**
M_{42}	**0.244**	0.137	0.158	**0.188**	**0.205**	**0.269**

that result in a translation, rotation and scale, invariant feature vector formation.

3.1.2 Experimental Work and Results

The performance of the proposed algorithm is evaluated by conducting the experiments are on a novel AbHA dataset and three publically available datasets - UR fall detection dataset, KARD dataset [78] and multi-view NUCLA dataset. K-Nearest Neighbour (K-NN) and Support Vector Machine (SVM) classifiers are used to classify the actions for the experiments. It is not necessary that features are always linearly separable. Therefore, to handle non-linearly separable action features, Radial Basis Function (RBF) kernel based SVM is used for classification. To optimise the performance of the non-linear SVM classifier penalty C, and gamma γ are optimised for each dataset that reduces overfitting. Average recognition accuracy is calculated to measure the performance of the algorithm, mathematically defined as Eq. (3.9).

$$\text{ARA} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \times 100\% \tag{3.9}$$

where TP is true positive, TN is true negative, FP is false positive, FN is false negative. Effectiveness of the proposed work is measured in terms of Average Recognition Accuracy (ARA), and compared with similar state-of-the-arts. The comparisons are relatively limited for UR Fall detection dataset [156] and KARD dataset [156] because few works are reported on these datasets.

3.1.2.1 UR Fall detection dataset

It is a RGB-D dataset, which was introduced by Kwolek et al. [156]. It comprises of 40 instances of ADL activities and 30 instances for fall for two viewpoints – front view and top view. In this work, only depth instances are chosen from the front view, and AESI are generated for 22 fall and 22 ADL activities for evaluation purpose. Sample images of the RGB/Depth Fall and ADL sequences are shown in Fig. 3.5. For UR fall dataset, performance of recognition system is optimised for $'K'$=3, while using K-NN classifier and Leave-One-Out Cross Validation (LOOCV) technique is used. For non-linear SVM (C, γ) parameters are set to (0.4, default) while using one-vs-one, and 5-fold cross-validation approach. Where penalty factor, C balances the trade-off between training error and margin maximization. For large value of C the training error will be reduced. However, excessive increase in the value of penalty factor may risk in losing the generalization properties of the classifier. One vs one classification is preferred over one vs all to ensure adequate ratio between testing and training samples of the dataset.

The experimental results are presented in Table 3.2. The ARA obtained for UR Fall dataset is compared with similar state-of-the-art techniques in Table 3.3. From Table 3.2 it is clearly evident that integration of Zernike moment and $\mathcal{R}$ -transform improves the recognition ability of the framework, by making it view insensitive. In Table 3.3, some of the state-of-the-arts [156] [100] [157] have achieved higher recognition accuracy than the proposed framework. This is so because these works utilised combination of RGB-D and accelerometric data to encrypt the actions. However, the work [156] has achieved only 90% accuracy while utilising only depth maps.

Table 3.2: ARA of the proposed work for UR Fall Detection Dataset

Activities	Action descriptor	Fall (%)	ADL (%)	ARA (%)
SVM (%)	$\mathcal{R}$ -Transform	94.6	93.2	93.9
	$\mathcal{R}$ -Transform + Zernike moments	95.5	95.5	95.5
K-NN (%)	$\mathcal{R}$ -Transform	94.7	96.28	95.89
	$\mathcal{R}$ -Transform + Zernike moments	96	97	**96.5**

Table 3.3: Comparison of ARA with other state-of-the-arts for UR fall detection dataset

Method	Classifier	Input	ARA (%)
Riemannian manifold [158]	SVM	RGB + v	96.77
V-DGP [156]	SVM	Depth maps	90
	SVM	RGB-D + Accelerometer	94.22
	K-NN	RGBD + Accelerometer	95.71
HTP [159]	SVM	RGB-D	87.76
EWMA [100]	SVM	RGB + Accelerometer	96.77
Curvelet [157]	SVM-HMM	RGB	96.88
OFFD [160]	CNN	RGB	95
Proposed Method	**K-NN**	**Depth maps**	**96.5**

3.1.2.2 KARD Dataset

Gaglio et al. [78] introduced Kinect Activity Recognition Dataset (KARD) in 2015. It is composed of eighteen activities as classified in Fig. 3.5. Ten different individuals perform each activity three times. Therefore, the dataset consists of $540\ (18 \times 3 \times 10)$ sequences captured at 640×480 resolution with 30 fps. For activity recognition K-NN (K = 5) and SVM with one vs one, 5-fold cross-validation technique is used. The performance of the proposed work on KARD dataset is provided in Fig. 3.5, which is 1.74% higher than the recent [161]. In Table 3.4, action recognition performance of the proposed work is compared with other state-of-the-arts exhibiting the generalisability of the proposed approach with superior accuracy.

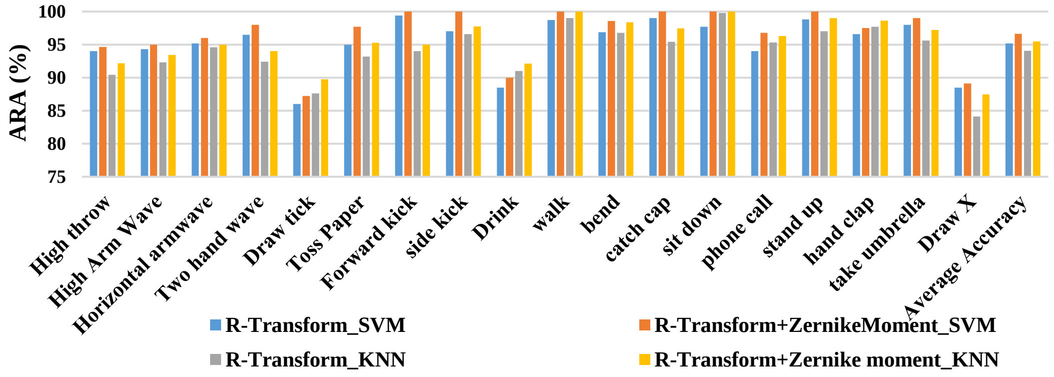


Figure 3.5: ARA for KARD dataset using SVM and K-NN

3.1.2.3 Abnormal Human Action (AbHA) Dataset

Abnormal Human Action Dataset (AbHA) [162] includes a novel set of commonly occurring abnormal actions in day-to-day lives of the elderly people such as ‘chest pain’, ‘headache’, ‘fainting’ and ‘falling backward’ and ‘falling forward’ as shown in Fig. 3.6. Abnormal actions, addressed here, stands for uncomfortable human postures in which the person require assistance in context of elderly people. Due to non-availability of the publicly available dataset for such activities, we generated our own dataset namely AbHA dataset. Which includes five actions, each performed by eight individuals and repeated two times. Hence, total generated samples are 80 $(8 \times 2 \times 5)$. Microsoft Kinect Depth sensor v1 is used to extract fine binary silhouettes by combing both skeleton joint coordinates over depth images. It simplifies the process of background subtraction to obtain binary silhouettes in real time scenario as the Kinect sensor is comparatively insensitive to the noise, clutter and illumination variations. Two third samples are used for training and one third for testing. The classification results on AbHA dataset are presented in Fig. 3.7. It is clear from Fig. 3.7 that the recognition rate for a headache and chest pain is 100 % because these actions are highly discriminating, whereas forward fall, fainting, backward fall, have substantial resemblance in their postures which resulted in lower recognition rate.

3.1.2.4 NUCLA Multi-View Action dataset

The NUCLA multi-view action dataset [165] is a collection of 10 actions performed by 10 subjects and acquired from three views: (i) left, (ii) front, and (iii) right, using Microsoft Kinect v1 as RGB-D videos . The actions of the dataset are (l) one hand pick up, (2) two hand pick up, (3) drop trash, (4) walk around, (5) sit down, (6) stand up, (7) donning, (8) doffing, (9) throw, and (l0) carry. The dataset possesses similar actions which makes it very challenging such “one hand pick up” and “two hand pick up”. For the experiments, two views are used for training, and rest for testing. It is observed from the Table 3.5 that the view variations of the actions are handled quite well with the integration of Zernike moment with $\boldsymbol{\mathcal{R}}$-transform, improving the performance of the framework by a significant amount 3.87% for K-NN($'K'$=5) and

0.6% for RBF-SVM classifier. Non-linear SVM is optimised for (C, γ) hyper-parameters as (0.99,default) values using one-vs-one, and 5-fold cross-validation. For K-NN classifier, LOOCV technique is used.

Table 3.4: Comparison of ARA with other state-of-the-arts for KARD dataset

Method	Classifier	Input Data	ARA (%)
Cippitelli et al. [161]	SVM	Skeleton	94.9
Gaglio et al. [78]	SVM	Skeleton	94.2
Madany et al. [163]	ConvNet	Skeleton	98.5
Pham et al. [164]	ResNet-44	Skeleton	99.97
Proposed method	**SVM**	**Depth**	**96.64**

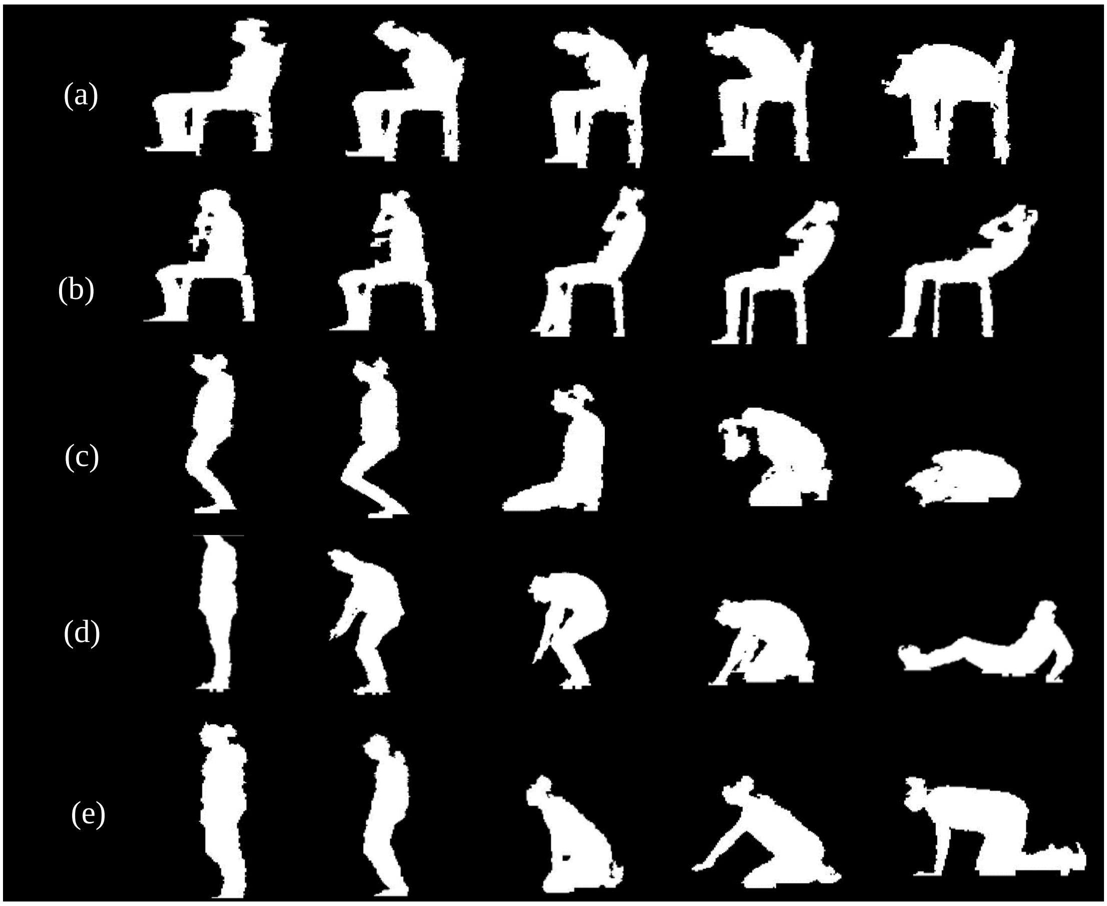
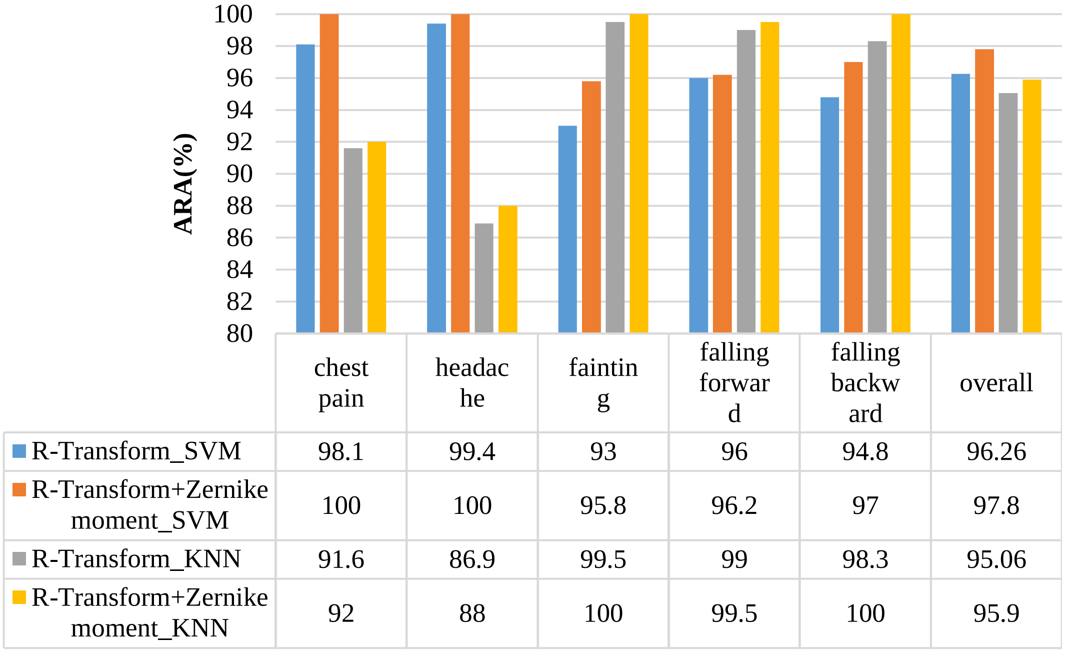


Figure 3.6: Binary silhouettes of AbHA dataset (a) chest pain (b) headache (c) fainting (d) falling backward (e) falling forward



	chest pain	headache	fainting	falling forward	falling backward	overall
■ R-Transform_SVM	98.1	99.4	93	96	94.8	96.26
■ R-Transform+Zernike moment_SVM	100	100	95.8	96.2	97	97.8
■ R-Transform_KNN	91.6	86.9	99.5	99	98.3	95.06
■ R-Transform+Zernike moment_KNN	92	88	100	99.5	100	95.9

Figure 3.7: ARA for AbHA dataset

Table 3.5: Comparison of ARA with other state-of-the-arts on multi-view NUCLA dataset

Method	**Input**	$\boldsymbol{V^{(3)}}$	$\boldsymbol{V^{(2)}}$	$\boldsymbol{V^{(1)}}$	**ARA (%)**
Depth-DVV [167]	Depth	58.5	55.2	39.3	51.0
CV-CVP [168]	Depth	60.6	55.8	39.5	52.0
NKTM [169]	RGB	75.8	73.3	59.1	69.4
R-NKTM [37]	RGB	78.1	-		78.1
HPM [166]	RGB-D	91.7	73.0	79.0	81.3
Skepxels [170]	Skeleton	91.5	85.5	79.2	85.4
Proposed method	$\boldsymbol{\mathcal{R}}$-transfrom_KNN	89	82.4	75	82.13
	Hybrid vector_KNN	91.8	86.1	80.1	86
	$\boldsymbol{\mathcal{R}}$-transform_SVM	90.6	85.6	79	85.06
	Hybrid vector_SVM	92.0	86.7	80.5	**86.4**

The real-time performance of the proposed work is verified on single NVIDIA GeForce 940M Graphics Card, Intel core i5 Processor and 8GB RAM. The processing time required to generate the feature vector and the time required to test the action, are computed, given in Table 3.6, that clearly proves the superiority in action recognition for the proposed framework in terms of processing and testing time than [166].

Table 3.6: Comparison of computation time of the proposed framework on multi-view NUCLA Dataset

Method	**Feature vector formation/ action (per video)**	**Testing time per sample (per video)**
$\text{HPM}_{\text{RGB-D}}$ GAN Refined Model [166]	49.1ms	0.68ms
Proposed framework	**0.95ms**	**0.33ms**

3.2 DoG-SDG based human action identification model

Understanding of human actions in videos is a growing field and received rapid importance due to surveillance, security, entertainment and personal logging. In this work, a new hybrid technique is proposed to describe RGB human action sequences. A unified framework endows a robust feature vector wrapping both global and local information that strengthens the discriminative depiction of action identification. For this purpose, initially, entropy-based texture segmentation is used for human silhouette extraction followed by construction of average energy silhouette images (AEIs). AEIs are the 2D binary projection of human silhouette frames of the video sequences, which reduces the feature vector generation time complexity. Spatial Distribution Gradients (SDGs) are computed at different levels of resolution of sub-images of AEI consisting overall shape variations of human silhouette during the activity. Scale, rotation and translation invariant properties of STIPs are used to develop a richer the vocabulary of DoG based STIPs using vector quantization which is unique for each class of the action. Experiments are performed to observe the behaviour of the proposed approach on four standard benchmarks i.e. Weizmann, KTH, Ballet Movements, Multi-view IXMAS. Promising results are obtained when compared with the similar state-of-the-arts, demonstrating the robustness of the proposed hybrid feature vector for different types of challenges –illumination, view variations posed by the datasets.

3.2.1 Proposed Methodology

We propose a novel hybrid technique for feature mining for human action recognition on standard HAR datasets. The uniqueness of this work is in the integration of AEI based spatial distribution gradients (SDGs) with scale, rotation and view-invariant Spatio-temporal interest points (STIPs). The hybrid feature is used to train SVM and HMM for action recognition and to compare the accuracy of the two classifiers. The block diagram depicts the flow of the proposed work in Fig. 3.8.

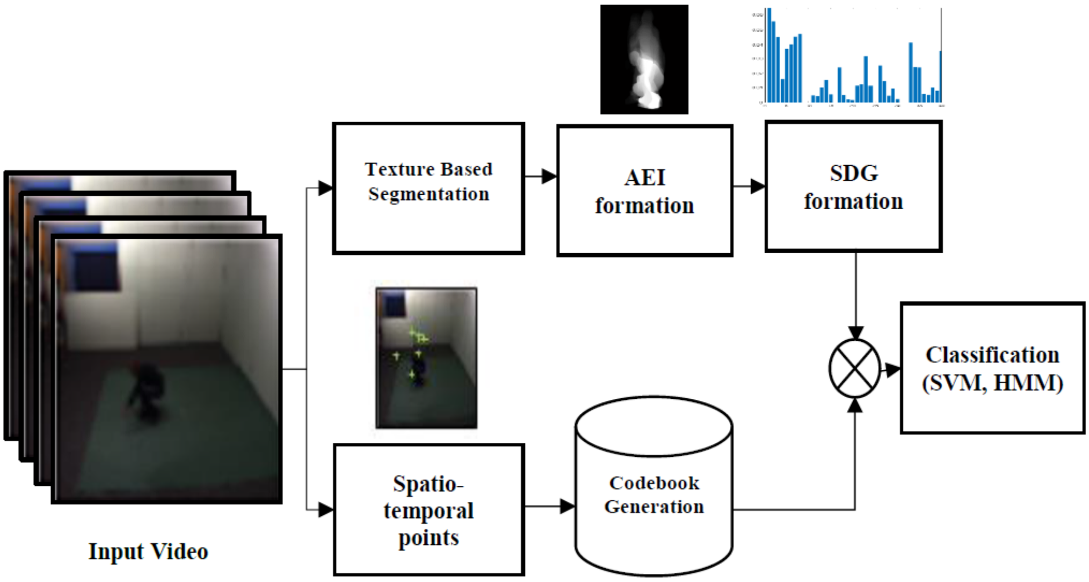


Figure 3.8: Flow diagram of the proposed framework

3.2.1.1 Spatial Distribution Gradients (SDGs)

A 2D spatial representation of an image includes a significant amount of information for non-verbal communication, which can be mined using spatial distribution descriptor of the human posture in the image. In this work, SDG is applied on Average Energy Images for each video sequence describing entire shape variations of the object in one image instead of accessing every frame or key frames of the videos.

Keyframe of videos are used in [171] [172] for feature generation, but it results in loss of temporal information, which can be preserved through AEI images.

3.2.1.1.1 Entropy based Texture Segmentation

For vision-based human action recognition, background extraction is an elementary objective. Addition of occlusion, background changes, illumination variations, and noise etc. [173] makes the task more challenging. An adaptive background extraction approach generates a background model and updates it frame by frame. If the frame possesses pixels which don’t satisfy background model, are treated as foreground pixels (human silhouette). In past years, Gaussian Mixture Model (GMM) [33] [34] and Local Binary Pattern (LBP) [174] [175] based strategies are widely used for texture-based foreground segmentation. The major problem of texture based segmentation approaches is that they are highly noise sensitive. Two objects, which need to be segmented, may have the same texture. To address these issues, Rampun et al. [176] defined Gray-Level Co-Occurrence Matrix (GLCM) based 32 features including eight Haralick’s statistical features [177]. Soh et al. [178] defined six features, which are Cluster Prominence, Dissimilarity, Entropy, Cluster Shade, Homogeneity and Maximum Probability.

Recently, implementation of GLCM based segmentation on FPGA [179] has provided a fast and efficient solution for real-time applications, which motivates us to use textural feature based segmentation technique using GLCM for human silhouette extraction. Entropy is the most broadly utilized parameter for depicting the textural properties. It is a factual measure of randomness in the gray level values of the image, mathematically given as Eq. (3.9):

$$Entropy = -\sum_i \sum_j \Omega(i,j)\, log\, \Omega(i,j) \qquad (3.9)$$

where $\Omega(i,j) = \frac{M(i,j)}{\sum_{i,j} M(i,j)}$ is the probability density function, and i and j are indices of Gray-Level Co-occurrence Matrix M. Large value of entropy indicates the complexity of the image is high. An entropy filter is applied on every pixel using 9×9 neighborhood pixels to generate texture mask. For small entropy kernel size i.e. 3×3 , 5×5 minute texture information are extracted as a noisy element and for larger kernel size 15×15 required texture information is filtered out. However, for 9×9 kernel size, relevant texture information is obtained without any noisy element, as observed in Fig. 3.9 (a). The obtained binary mask is mapped with the original image to extract human silhouette Fig. 3.9 (b).

3.2.1.1.2 Average Energy Image (AEI) Computation

There are various features, which are used to represent human actions such as a bag of features, local descriptors, and global descriptors. In this work, average energy image is exploited to represent human actions. The concept of AEI is an extension of GEI which contains space and temporal features. It reduces the effect of the rate at which action is performed, hence reduces the intra-class variations. It is computed as the average sum of binary silhouette frames in a video. Let $\{ A_1, A_2, \dots \dots, A_n\}$ be the binary silhouettes of a video sequence performing an activity and n represents the frame number. Thus a set of binary silhouette images per action is

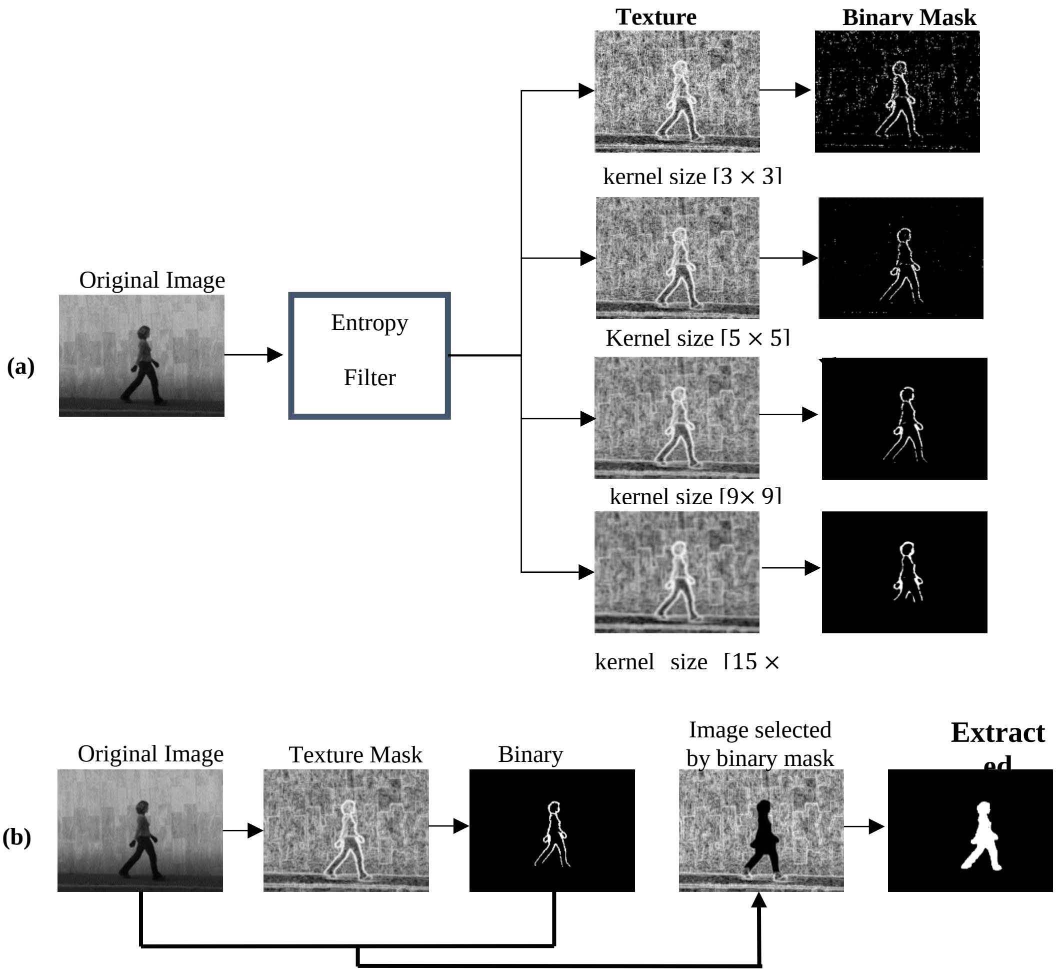


Figure 3.9: Entropy-based silhouette extraction (a) Entropy Filter Kernel (b) Extraction of a human silhouette

represented by one average energy image. It handles the problem of huge data storage and computational complexity involved in it. The average energy image (AEI) can be generated by using Eq. (3.10).

$$AEI(x,y) = \frac{1}{n}\sum_{i=1}^{n}|A_i(x,y)|^2 \qquad (3.10)$$

where x and y are coordinates of a binary image A_i. AEI in Fig. 3.10 shows, the AEIs of and ROIs of a video clip of an action. The bright section in AEI image Fig. 3.10 (c) and (d) represent the static portion of the body and less brightened section shows variations of the body during the activity. During the activity, the main body of the actor performing the hand-waving action is still, i.e. only hand movement is present.

Hence the middle portion of the curve represents the pixel's values at those points which are not in motion. The still body points have pixel intensity of value 1, whereas the hands have different pixels' values ranging from greater than to 0 to less than 1. To process a small region of interest is computationally more efficient than the whole image. Therefore, to further process the AEI, firstly ROI is extracted from the AEI frame by scanning the image from left to right column wise. The first column with one or more non-zero pixels is considered as ROI extreme left and the last column with non- zero pixel/s is considered as ROI extreme right. AEI image is accessed from top to bottom and first and last rows with non-zero pixels are taken as ROI top most and lowest extreme of the ROI. And a rectangle is drawn using these extreme points.

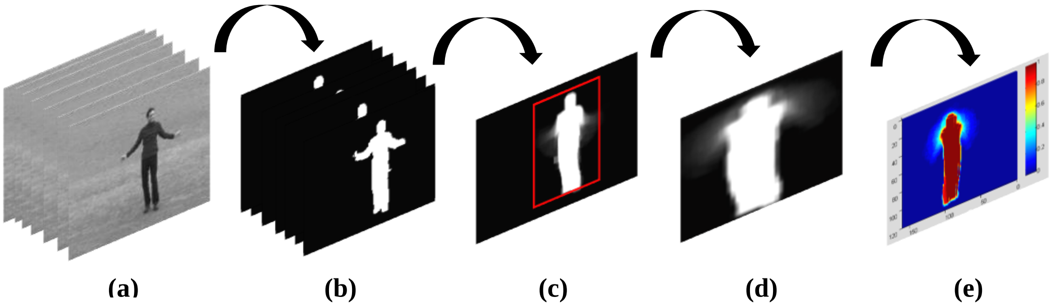


Figure 3.10: Illustration of AEI formation and ROI Extraction: (a) Input video (b) Entropy based silhouette segmentation (c) AEI formation (d) cropped AEI (e) AEI-3D Plot

In the experiments, the dimension of ROI 64 × 38 depends on the height and width of the person performing the action. To maintain the uniformity, the dimension of ROI of all the samples is resized to 64 × 38. Fig. 3.11 illustrates 'check watch' action at five different camera views used in IXMAs dataset and from these images the Spatio temporal variations are reflected in terms of intensity variations.

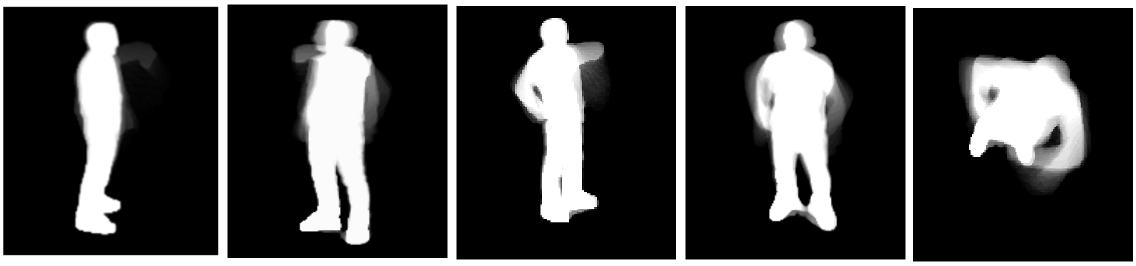


Figure 3.11: AEI image of 'check watch' action class of IXMAS datasets at different camera angles

3.2.1.1.3 Spatial Distribution of Gradients (SDGs) Computation

Spatial Distribution of Gradients (SDGs) [172] is a shape appearance-based object descriptor. In this work, SDGs descriptor is computed for AEIs of each human action video. This method aims to represent the shape of the object by providing a spatial distribution computational model. It divides the gradient of AEI $E(x, y)$ into 4^{ℓ} sub-regions at each level ℓ $\{0, 1, 2\}$. Increase in number of levels, increases the length of SDG feature vector and the maximum value of the SDG magnitude decreases as shown in Fig. 3.12. And the variation in SDG magnitude are quite comparable for ℓ=2 & 3 with respect to the maxima and minima of magnitude. Therefore, higher level decomposition will only increase the dimension of the descriptor, without any significant improvement in performance. The Spatial Distribution Gradient algorithm is explained in Algorithm 1 by considering 0, 1 and 2 level of decomposition. For each sub-region at level ℓ of decomposition, angle, and magnitude of Gradients $E(x, y)$ of AEI image are computed. Gradients represent the edges of the shape variations in AEI. The canny edge detector is used to obtain edges jointly. SDGs descriptor quantise the magnitudes of the AEI for 8-evenly spaced orientation bins which adds to dimensionality reduction and generates the histogram spatial gradient feature for each sub-region of the image at different levels. To guarantee that the image having a larger number of edges are not preferred over other images, SDG descriptor is standardized by normalization to unity. The vectors at every level are connected such that the loss of spatial data is reduced. The SDGs descriptor is analogous to the pyramidal outline of HOG descriptor. From Fig. 3.13 it can be observed that the computed SDG descriptor is insensitive to translation and scale variations but changes with rotation. Since, SDG descriptor is applied on ROI extracted from Average Energy Image (AEI), in the algorithm. ROI of the average energy remains unaffected by scale and translation variations. Thus, SDG descriptor remains invariant to translation and scale disparity. However, in case of rotation, due to deviation in ROI of an action, SDG descriptor magnitude distribution per bin changes.

3.2.1.2 Spatio Temporal Interest Points (STIP)

For action recognition, extraction of appropriate features is a critical task. Recently, STIPs have emerged as a popular means of local descriptor-based action recognition. However, distribution of STIPs should be stable around the object. A point in space and time is considered as a Spatiotemporal interest point (STIP) if it possesses prominent intensity variation in space as well as time. Intensity variation in space implies enormous contrast variation. Whereas, intensity variation in time domain occurs if a point varies over time significantly. In real-time surveillance applications, the person under observation can be captured from completely different camera viewpoints, with different scene compositions and resolution and scales. This introduces large intra-class separation among features, resulting in misclassification.

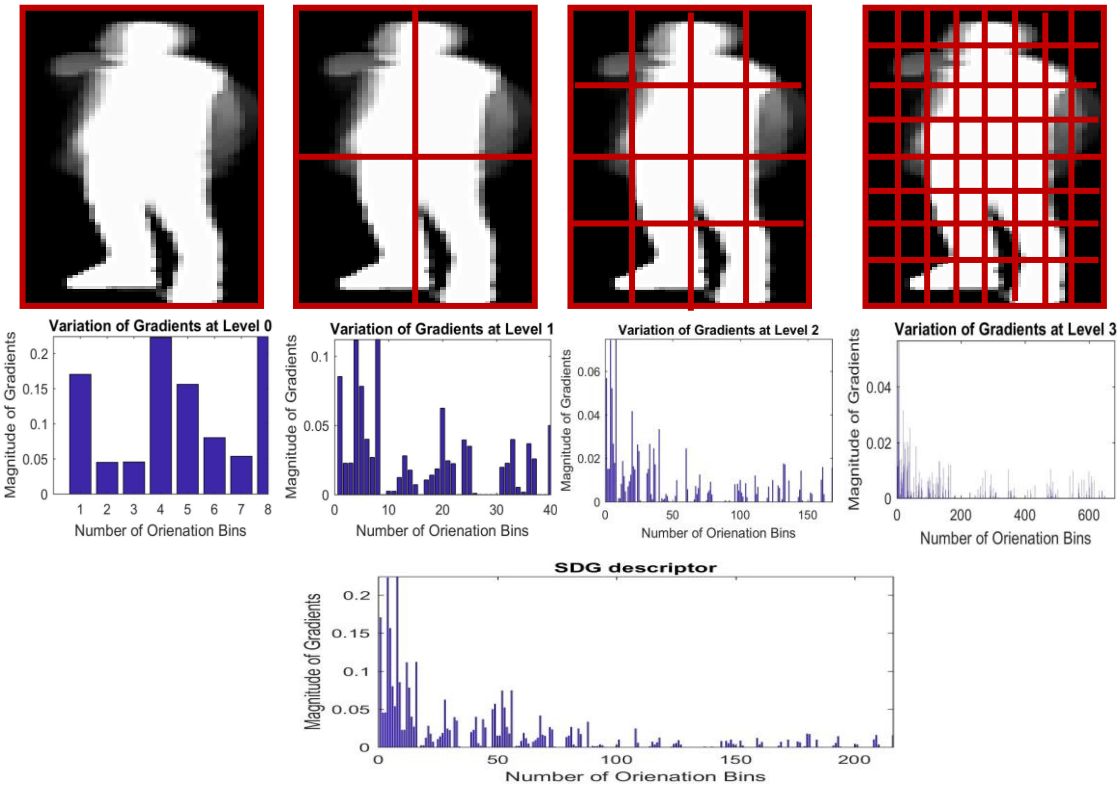


Figure 3.12: Spatial Distribution Gradients representation with level compositions

To handle the issue of scale variations scale, invariant STIPs are computed for video sequences by exploiting Difference of Gaussians (DoG).

3.2.1.2.1 Difference of Gaussian (DoG)

The Difference of Gaussian (DoG) is widely used to derive scale-invariant interest points. The incentive of using DoG is that it removes high-frequency spatial details

```
Algorithm: Computation of SDGs

Step 1: Input $AEI(i,j)$ of a video sequence $V(i,j,t)$.

Step 2: Apply Canny Edge Detector as $E(x,y) = Canny(AEI(i,j))$

Step 3: Compute SDGs at different levels as follows:

        (a) At level 0: Compute magnitude $\mathcal{M}(x,y)$ and orientation $\Theta(x,y)$ of
            $E(x,y)$ at any point $(x,y)$ using following formulas.
```

$\mathcal{M}(x,y) = [E_{Gx}(x,y)\ ^2 + E_{Gy}(x,y)\ ^2]\ ^{\frac{1}{2}}; \Theta(x,y) = arctan(\frac{E_{Gy}(x,y)}{E_{Gx}(x,y)})$

where $E_{Gx}(x,y)$ and $E_{Gy}(x,y)$ are image gradients of $E(x,y)$ image in x and y directions respectively. The magnitude values are quantized into K-evenly spaced orientation bins from 0^0 to 180^0. Hence, length of the generated histogram h_0 is computed as $L_\ell = K \sum_{\ell=0}^{2} 4^\ell$ where K=8 and $\ell = 0$ i.e. $L_0 = 8 \times 1 = 8$.

(b) At level 1: Image $E(x,y)$ is divided in four sub-regions $\{E_1(x,y), E_2(x,y), E_3(x,y), E_4(x,y)\}$ and a feature vector is framed by applying *step 4(a)* on each sub-region. Length of the generated histogram h_1 is $L_1 = 8 \times [1+4] = 40$.

(c) At level 2: Each sub-region $E_i(x,y), i = 1,2,3,4$ is further divided in four sub regions as $E_{ij}(x,y), i = 1\ to 4\ \ \&\ j = 1\ to\ 4$. Histogram h_2 is generated for each sub-region using step 4(a) with length $L_2 =$ $8 \times [1+4+16]$ =168.

Step 4: ***Output:*** Histogram of Spatial Distribution Gradient for an $AEI(x,y)$ is obtained as $h = \{h_0\ , h_1\ , h_2\}$

such as random noise, which is a common element in real-time applications. According to scale-space theory, each frame of the video is given multi- scale signal representation (x,y,σ) , mathematically defined using Eq. (3.11):

$$S(x,y,\sigma) = G(x,y,\sigma) * I(x,y) \tag{3.11}$$

where $G(x,y,\sigma)$ correspond to a Gaussian kernel function with kernel size $[25 \times 25]$, $(x,y) \epsilon\ R^{m \times n}$, $m \times n$ is the dimension of the image $I(x,y)$. A $[25 \times 25]$ Gaussian kernel captures strong DoG keypoints with 118 as highest DoG keypoint pixel value. For smaller Gaussian kernels i.e. $[9 \times 9]$, $[3 \times 3]$ the highest DoG keypoint pixel value obtained are 104 and 65 respectively.

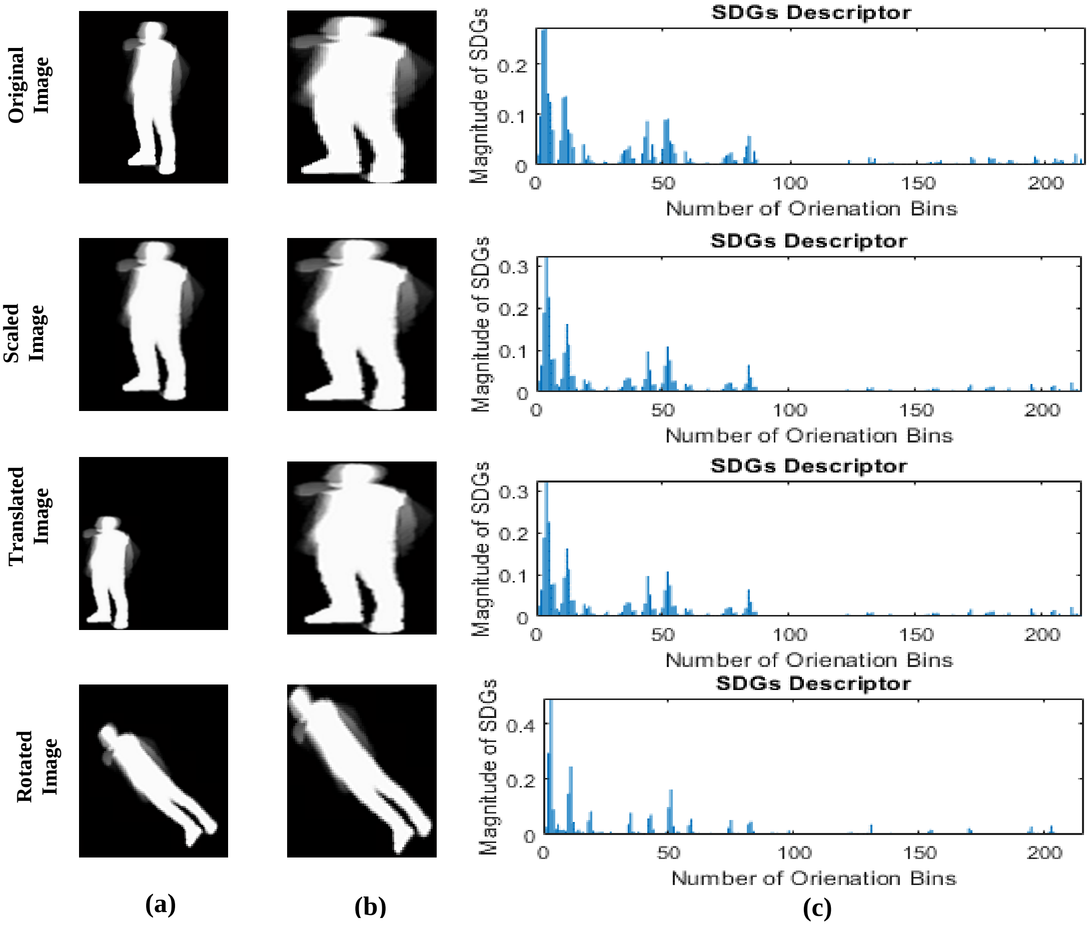


Figure 3.13: Geometric invariance of SDG descriptor (a) input image (b) extracted ROI (c) SDG descriptor

As the kernel size is increased further $[50 \times 50]$ the possible highest DoG keypoint pixel value saturated to 118. Therefore, the Gaussian kernel $[25 \times 25]$ is selected for DoG based key features extraction. Mathematically $G(x, y, \sigma)$ is defined as Eq. 3.12:

$$G(x, y, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{3.12}$$

For the identification of stable key-points or edges in scale space, $DoG(x, y, \sigma)$ is defined as:

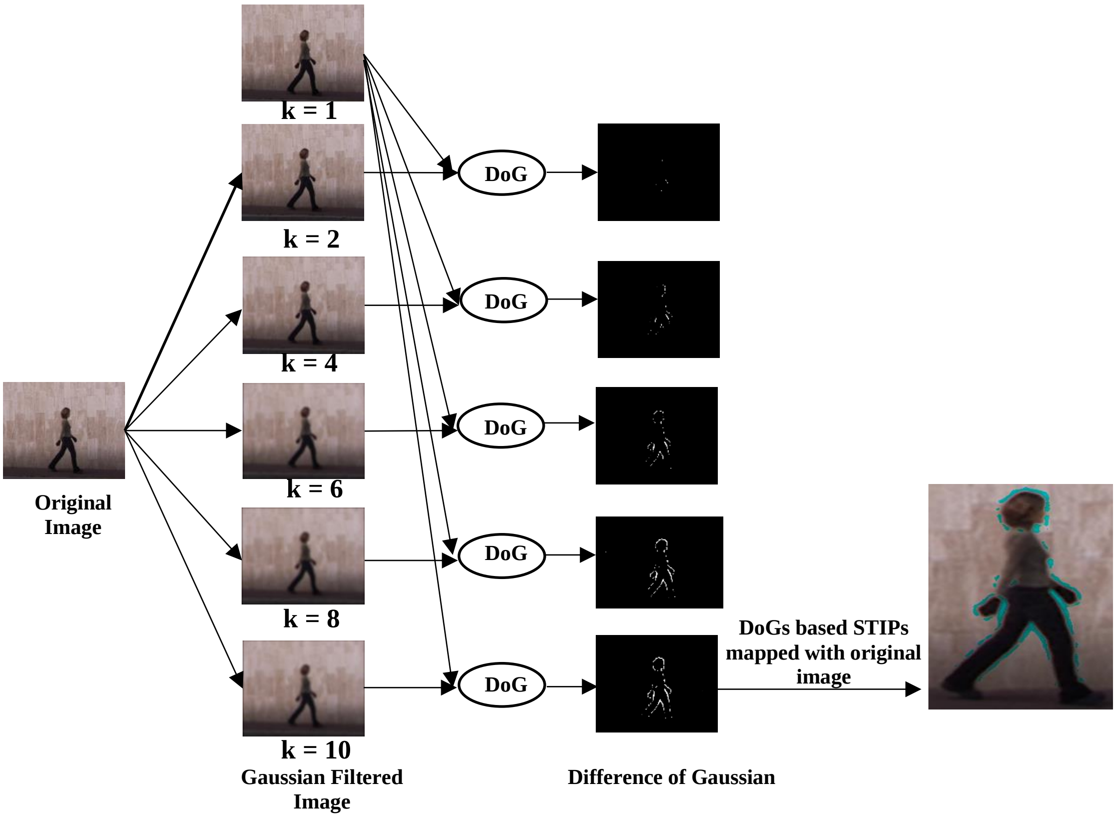


Figure 3.14: Extraction of. DoG based STIPs, k = scale of Gaussian function

$$DoG(x,y) = S(x,y,k\sigma) - S(x,y,\sigma) \;=\; [G(x,y,k\sigma) - G(x,y,\sigma)] * I(x,y) \quad (3.13)$$

For experimental work, σ is chosen to be 0.7 and k is the scale parameter chosen to be 10, to optimise the prominent key features. High value of the Difference of Gaussian pixel represents the strong interest points which are independent of scale variations. The obtained $DoG(x,y)$ is passed through a filter with the threshold value as 10% of the highest DoG pixel value per frame. It rejects the pixels lesser than the threshold value. The obtained key-points per frame are concatenated to maintain the codebook $[\, x_k \; y_k \; p_k]$, $k\, = 1$: no of key points per frame for each video sequence.

3.2.1.2.2 Codebook Generation

A vocabulary of Spatio-Temporal key interest points is created for better action representation. For each frame in the video, there is a key point vector, which depicts the number of interest points in that frame. Usually, there are two kinds of approaches [180] for codebook generation algorithm i) partition each feature vector space represented by its centre called code-word ii) to compute the probability distribution of features using a generative model such as GMM. Under category one, there are many vector quantization approaches such as K-means clustering [180], hierarchal clustering [181] and spectral clustering [182]. Among them, K-means clustering is the widely used approach to construct codebook. For a set of local features $\{q_1, q_2, \dots q_n\}$ where $q_n \in \mathbb{R}^D$our objective is to partition the local feature vector in K clusters as $\{f_1, f_2, \dots f_K\}$, where $f_K \in \mathbb{R}^D$. For each feature q_n a binary indicator variable $b_{nK} \in \{0,1\}$ is assigned. If q_n is allotted to K^{th}cluster $b_{nK} = 1$ and $b_{ni} = 0$ if $i \neq K$. The objective function is defined as: $\min \zeta(\{b_{nk}, f_k\}) = \sum_{n=1}^{N} \sum_{k=1}^{K} b_{nK} \|q_n - f_k\|^2$. The values of $\{b_{nk}\} and\ \{f_k\}$ are optimised to minimize the objective function ζ iteratively. The fundamental steps involved in codebook generation are illustrated in Fig. 3.15. The extracted DoG based STIPs are stored as $\{x_k, y_k, p_k\}$, $k =$ no of STIPs per frame, $\{x_k, y_k\}$ is the x and y coordinates of the k^{th}STIP and p_k is the pixel value of the k^{th} STIP. Hence a feature vector of dimension $3 \times k$ is formed per frame. For a video sequence of n no. of frames, a feature vector of dimension $3 \times k \times n$ is constructed. It is quantised using K-means clustering approach [180], where K is optimised with K=128 for maximized recognition accuracy for the dataset. The optimised feature vector with dimension (3×128)' for each video is collected in the codebook of dimension $128 \times 3 \times m$.

The final feature vector is designed by integrating SDGs and Codebook representation of extracted DoG based Spatio-Temporal interest points for classification. This fusion model is enriched with shape and motion evidence and has the power of recognizing the activity distinctly. In Fig. 3.16, histogram patterns of the designed hybrid feature vector are illustrated for KTH dataset activities, which depicts unique patterns for each activity. This property of hybrid feature vector assures large inter-class separation.

3.2.2 Experimental Work and Results

To observe the performance of the proposed framework four publicly available datasets: KTH [183], Weizmann [184], Ballet [185] and IXMAS [186] dataset are used. It helped to verify the strength of the proposed approach against the illumination change, viewpoint variation, high interclass similarity, and low intra-class similarity of the actions. The image samples for each dataset are provided in Fig. 3.17. The performance of the algorithm is measured in terms of Average Recognition Accuracy (ARA) using HMM and multi-class SVM classifier. For experimentation, the standard HMM is defined for n no. of output states (no. of action classes) and tested for 2 to 11 no. of hidden states. Expectation Maximisation (EM) algorithm is used to estimate the model parameters for 80 iterations.

3.2.2.1 Weizmann Dataset

Weizmann dataset [184] consists of total 90 video sequences with 25fps frame rate and 144×180 frame size. The dataset includes 10 different actions perfromed by 9 people, as shown in Fig.3.19 (a).

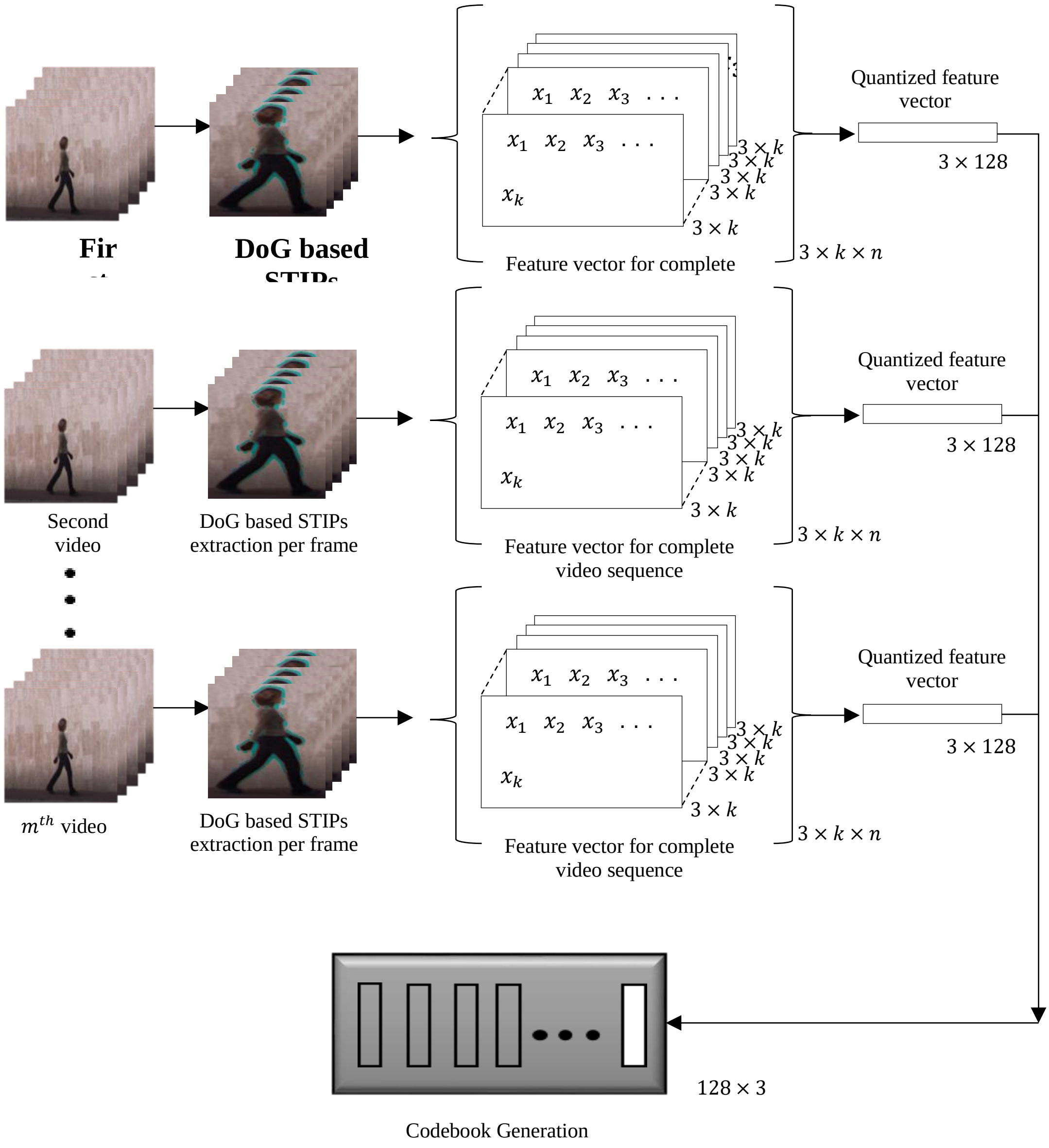


Figure 3.15: Illustration of codebook generation of STIP feature vector

3.2.2.2 KTH Dataset

The KTH dataset [183] consists 100 videos sequences for 6 basic activities, as shown in Fig. 3.19 (b), in indoor and outdoor. These video sequences are acquired in the constant background condition with a fixed calibrated camera, and 25fps frame rate with 160×120 frame size. It is a larger and more challenging dataset than Weizmann dataset.

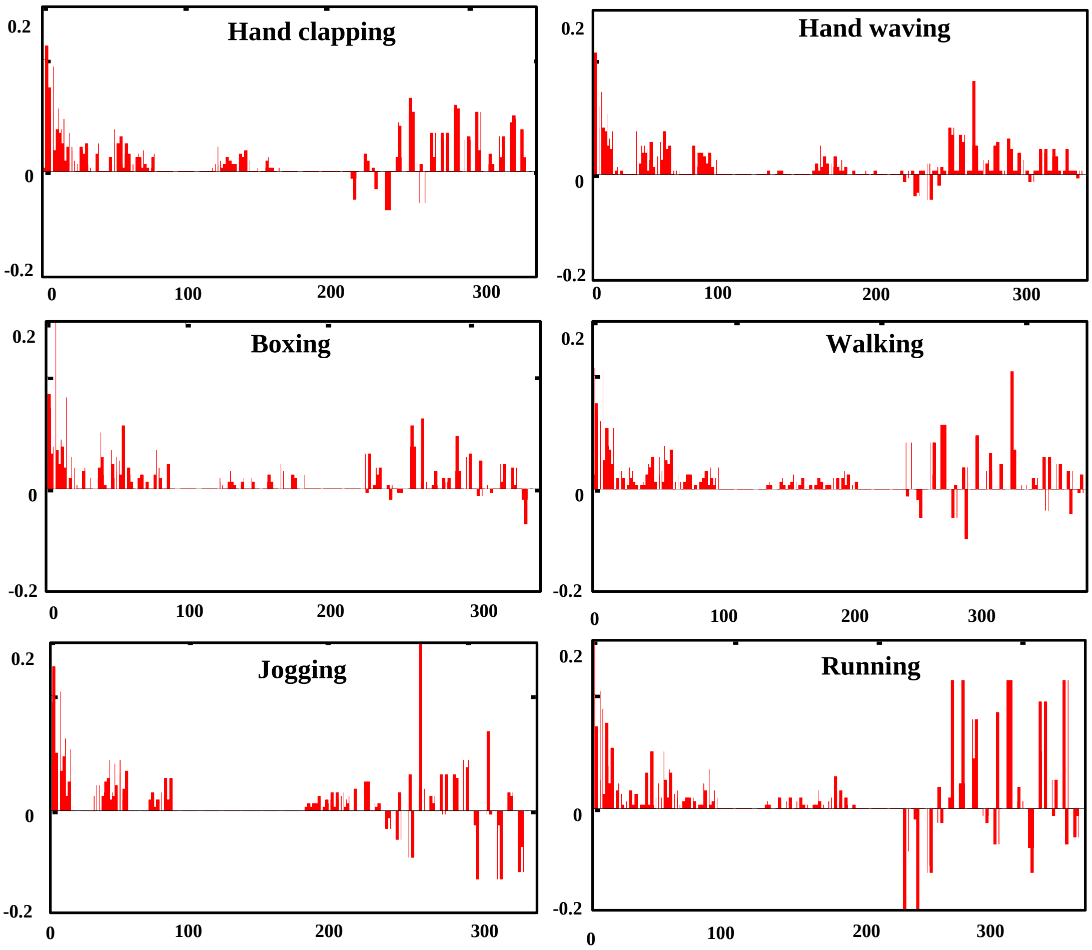


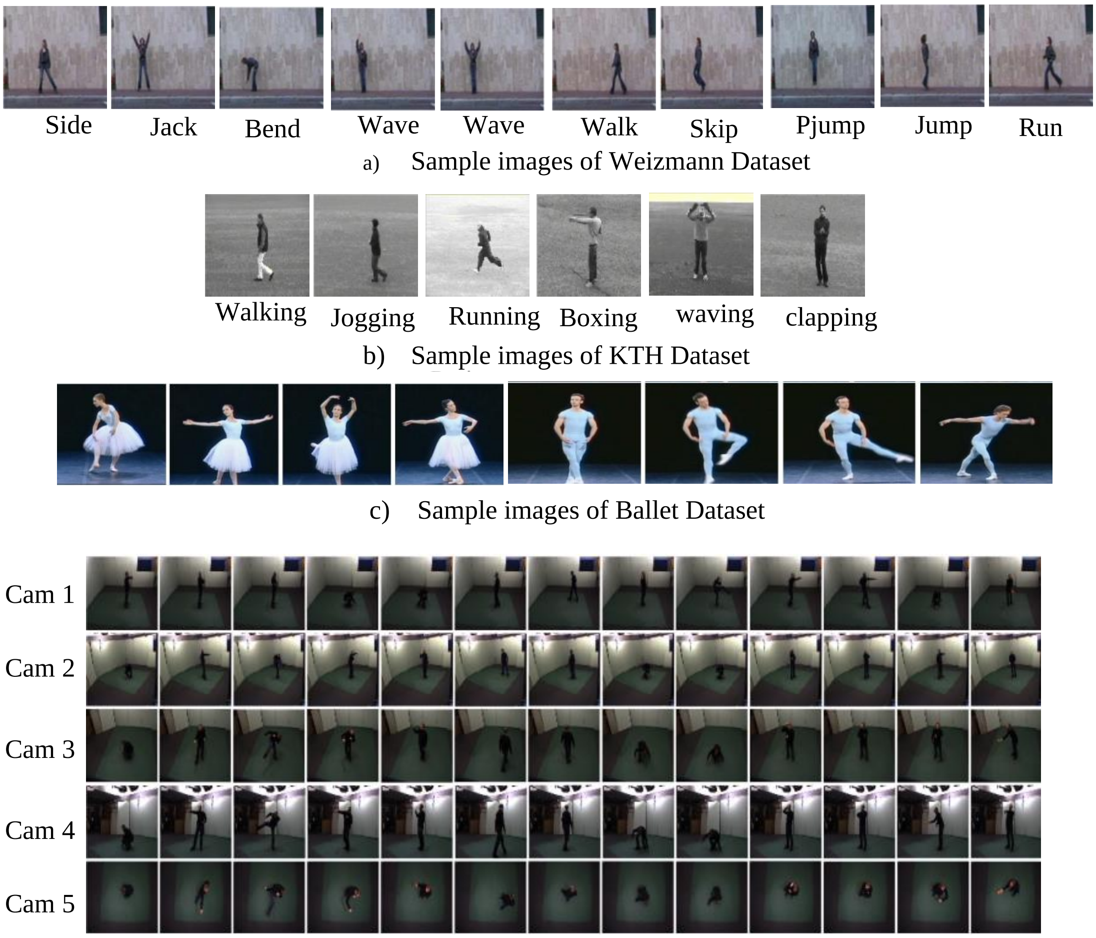
Figure 3.16: Histogram pattern of Hybrid feature vectors of various KTH dataset activities

3.2.2.3 Ballet Dataset

Ballet dataset [185] consists of eight human ballet movements, as shown in Fig. 3.19 (c). It offers significant amount of intra-class dissimilarity and inter-class similarity in terms of scale variations, speed of the action performed, and clothes.

3.2.2.4 IXMAS Dataset

INRIA Xmas Motion Acquisition Sequences (IXMAS) dataset [186] is a view-invariant human action recognition dataset. It is one of the widely used dataset to analyse human actions performed under different views. It includes 13 daily living activities recorded via five cameras with $23fps$ frame rate, which are shown in Fig.3.19 (d). These actions are enacted by 12 actors repeated three times with 390×291 spatial resolutions. The dataset offers the challenges in terms of variation in clothing, sex, execution rate of the actions and different heights of the actor preforming the action. The performance of the proposed approach is evaluated in terms of ARA on four different datasets, as shown in Table 3.7.



a) Sample images of Weizmann Dataset

b) Sample images of KTH Dataset

c) Sample images of Ballet Dataset

d) Sample images of IXMAS Dataset for five different

Figure 3.17: Sample frames a) Weizmann b) KTH c) Ballet d) IXMAS human action dataset

The performance of the proposed algorithm is compared with Human Pose Model (HPM) and Human Pose Model-Temporal Modelling (HPM+TM) [187] for each dataset. Due to non-availability of a large and sufficient set of poses, Rahmani et al. [187] used CMU Motion Capture Database [188] which covers over 2600 mocap sequences trained for synthetic human poses in different views from 0^0 to 180^0 . It supports the view invariance property of the HPM architecture.

Table 3.7: **Average Recognition Accuracy (ARA) of the proposed algorithm using SVM and HMM classifier**

Classifier/ Dataset	**ARA(%) with HMM**	**ARA(%) with SVM**
Weizmann	96.2	**97.5**
KTH	95.8	**96.6**
Ballet	94.3	**95.62**
IXMAS	88.36	**89.18**

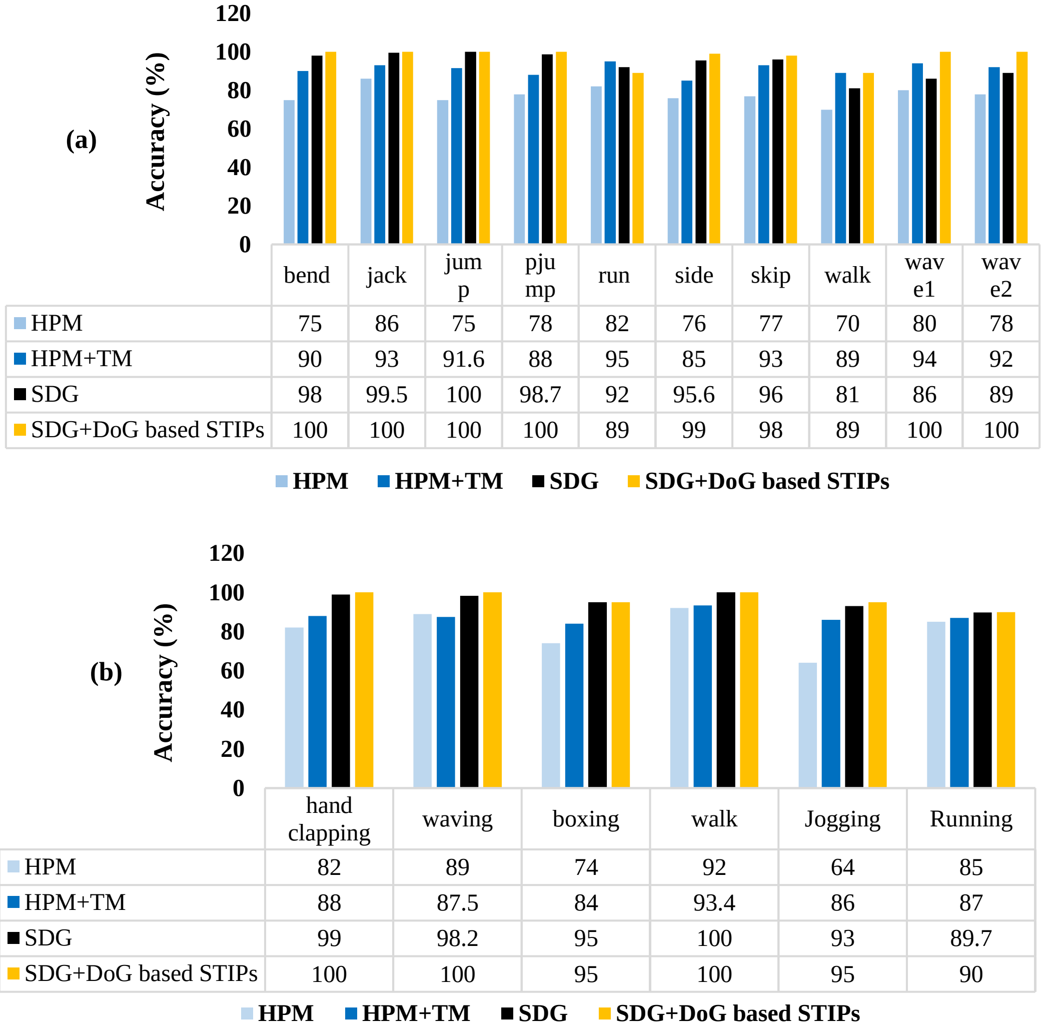
It motivated us to compare the adaptive capacity of the proposed work against view changes. HPM is trained for depth sequences. Though, selected datasets, in the work, possess only RGB frames. Therefore, binary silhouettes are fed to HPM model instead of depth silhouettes. The concept of transfer learning is used to fine tune the HPM model for 10 classes of Weizmann dataset, 6 classes for KTH, 7 classes for ballet and 13 classes for IXMAS dataset, which describe each frame as feature vector 4096×1 . Fourier Temporal Pyramid (FTP) feature vector 4096×28 is computed to encode the temporal details of $4096 \times n$ HPM feature vector S_i for i^{th} action sequence with a number of frames in the action sequence, using a pyramid of three levels which divides $S_i$in equal halves at each level as 1 + 2 + 4 = 7 feature groups. Short Fourier Transform is applied to each group. It generates spatio-temporal action descriptor in the form of four low-frequency coefficients $(4 \times 7 = 28)$. It is noticed that temporal encoding has definitely improved the recognition accuracy than HPM [187]. However, significantly higher accuracy is obtained for the proposed the proposed work. The confusion matrix for the same is shown in Fig. 3.19(a)-(d). For Weizmann dataset state-of-the-art with ARA of 97.5% is achieved. Despite of clothing variation present in

Weizmann dataset, texture based segmentation for silhouette extraction is accountable for such high recognition rate. Integration of SDG and DoG based STIPs feature vectors achieves 100% recognition accuracies for ‘bend’, ‘jack’, ‘jump’, ‘pjump’, ‘wave1’ and ‘wave2’, as shown in Fig. 3.18(a). Though in KTH dataset there are smaller number of action classes, the recording conditions are more irregular than Weizmann. Hence, KTH is more challenging dataset than Weizmann dataset due to different setups and scale variations. Therefore, silhouette extraction is an important and challenging task for KTH dataset. Background subtraction approach is also sensitive to illumination and recording conditions’ variations. Whereas, texture based foreground extraction approach makes it insensitive to illumination and recording conditions. In our experiment, the highest ARA achieved using texture based silhouette extraction methodology for human activity recognition is 96.6% for KTH dataset. Proposed algorithm (SDG + DoG based STIPs) evaluation results on KTH dataset are given in Fig. 3.18(b) and also compared with HPM and HPM+TM method [189].

In Ballet dataset size variation, clothing, sex and execution rate of actions introduced additional complexity that makes action recognition even more challenging. The proposed hybrid feature design is insensitive to execution rate and size variation of the actor while performing the action. Because it incorporates Difference of Gaussian (DoG) based scale invariant Spatio-temporal key points (STIPs) and AEI based Spatial Distribution Gradients (SDGs) which is insensitive to the speed of action.

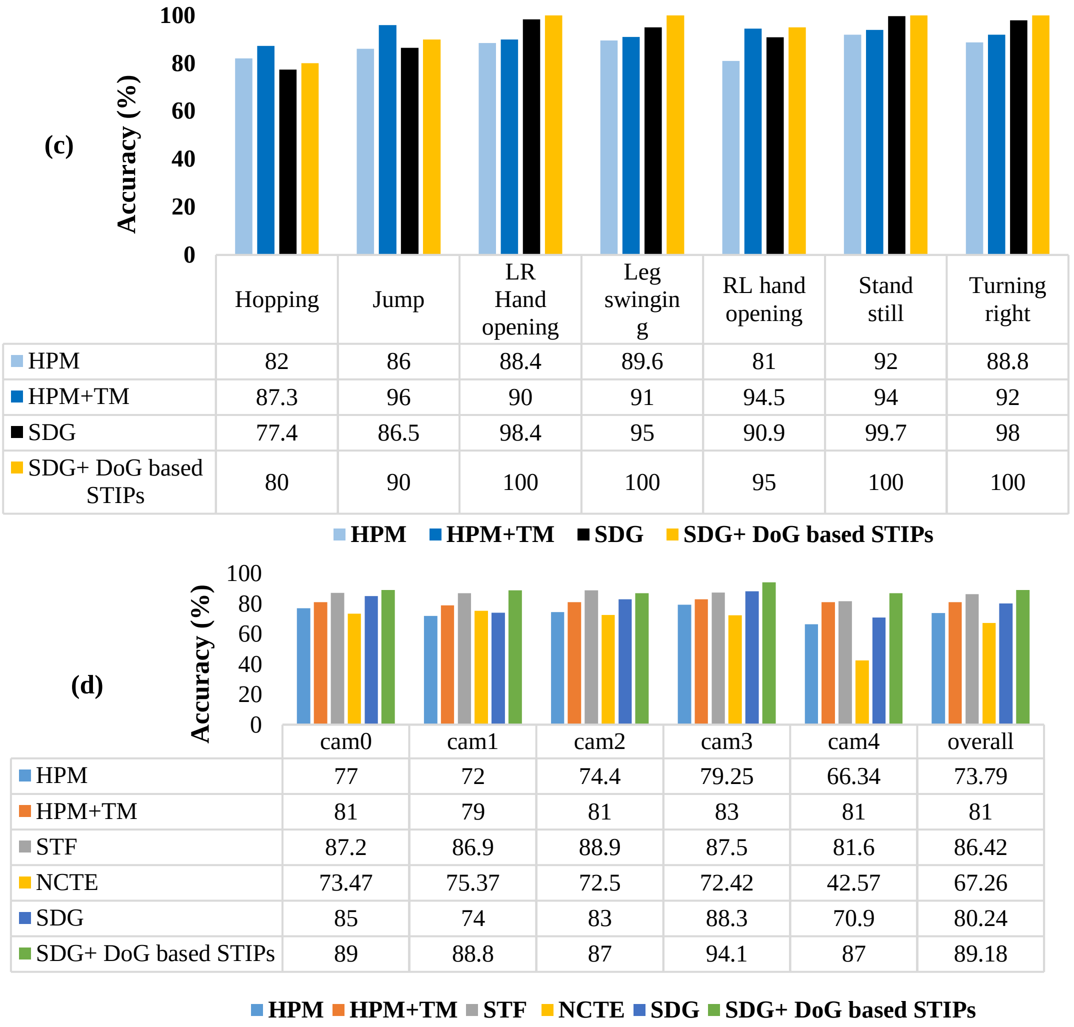
For AEI formation sequence of human pose energy is collected irrespective of the speed of action performed. Fusion of DoG based STIPS with SDG feature vector has increased recognition accuracy from 92.2% to 95.62% ARA for Ballet dataset, Fig.

3.18 (c). From confusion matrix, in Fig. 3.19(c), it is quite evident that except ‘hopping and jumping’ misclassification, appreciable recognition results are obtained for all the activities in the dataset despite of the present complexity, as mentioned above. Table 3.8 reports the ARA evaluated for IXMAS dataset by the proposed algorithm. For experiment, pair-wise train-test camera views are selected and classification is performed.



	bend	jack	jum p	pju mp	run	side	skip	walk	wav e1	wav e2
■ HPM	75	86	75	78	82	76	77	70	80	78
■ HPM+TM	90	93	91.6	88	95	85	93	89	94	92
■ SDG	98	99.5	100	98.7	92	95.6	96	81	86	89
■ SDG+DoG based STIPs	100	100	100	100	89	99	98	89	100	100

	hand clapping	waving	boxing	walk	Jogging	Running
■ HPM	82	89	74	92	64	85
■ HPM+TM	88	87.5	84	93.4	86	87
■ SDG	99	98.2	95	100	93	89.7
■ SDG+DoG based STIPs	100	100	95	100	95	90



	Hopping	Jump	LR Hand opening	Leg swingin g	RL hand opening	Stand still	Turning right
■ HPM	82	86	88.4	89.6	81	92	88.8
■ HPM+TM	87.3	96	90	91	94.5	94	92
■ SDG	77.4	86.5	98.4	95	90.9	99.7	98
■ SDG+ DoG based STIPs	80	90	100	100	95	100	100

■ HPM ■ HPM+TM ■ SDG ■ SDG+ DoG based STIPs

	cam0	cam1	cam2	cam3	cam4	overall
■ HPM	77	72	74.4	79.25	66.34	73.79
■ HPM+TM	81	79	81	83	81	81
■ STF	87.2	86.9	88.9	87.5	81.6	86.42
■ NCTE	73.47	75.37	72.5	72.42	42.57	67.26
■ SDG	85	74	83	88.3	70.9	80.24
■ SDG+ DoG based STIPs	89	88.8	87	94.1	87	89.18

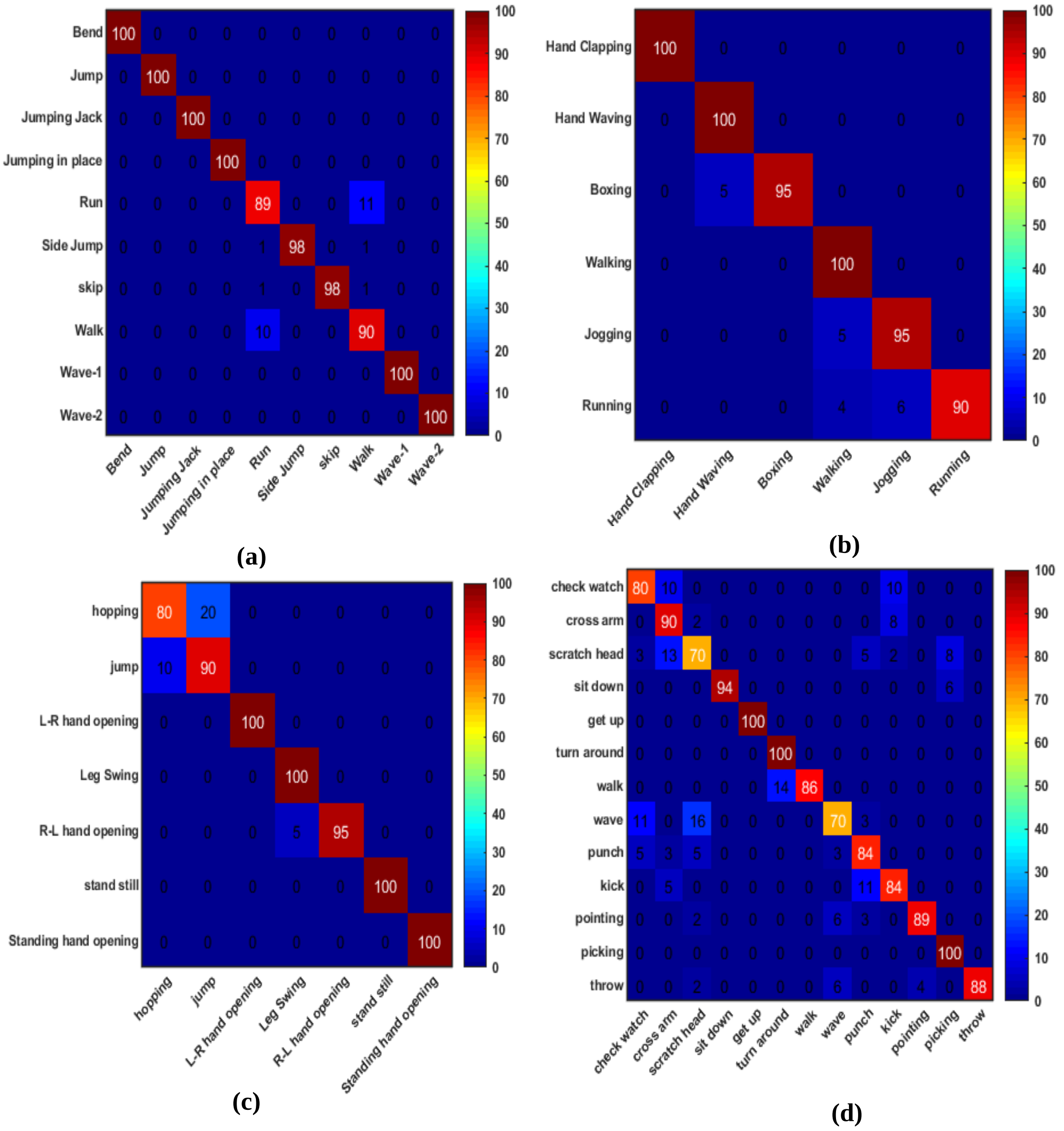
■ HPM ■ HPM+TM ■ STF ■ NCTE ■ SDG ■ SDG+ DoG based STIPs

Figure 3.18: Per class performance of the proposed method (SDG + DoG based STIPs) vs HPM and HPM+TM [176] for (a) Weizman dataset (b) KTH dataset (c) Ballet dataset (d) Multi-view IXMAS dataset

In First row in Table 3.8 indicates test view and first column shows selected training view. Complexity of viewpoint variations is very well handled by simple yet effective DoG based STIPs feature vector which has increased the overall recognition accuracy from 80.24% to 89.18%, when fused with SDG descriptor. Fig. 3.19(d) depicts the confusion matrix for IXMAs dataset with 13 activities when cam2 is used for testing and (train-test) pair-wise multi-view testing is performed. The performance of the proposed algorithm is also compared with other state-of-arts, as presented in Tables 3.9 to 3.12. It is evident from these comparisons, that the obtained recognition accuracy of the proposed method is superior from other approaches. Hence, the proposed framework is proved to be more robust than other state-of arts for human activity recognition, tested for variable conditions offered by the datasets. There are number of approaches [190] [191] [192] [193] [194] [195] [196], which used Weizmann dataset to evaluate the algorithm but only few achieves comparable accuracy such as BoCP [191] and VCHA [196], Table 3. BoCP [191] unified bag of correlated silhouette poses and MHI for local and global action description. A bioinspired computational model [196] received little higher recognition accuracy than ours as which recognized human actions by stimulating computationally intensive neural networks. A comparative study of the performance of the proposed work with other state-of-the-arts on Ballet dataset is shown in Table 3.11. It is one of the toughest dataset in terms of human action's complexity. Since, the actions are performed in in controlled environment, silhouette extraction process generated satisfactory results for the dataset. The proposed feature design achieved, slightly higher ARA than Vishwakarma et al. [58] due to scale, illumination, translation and view-invariant property of the hybrid feature, which is the integration of DoG based STIPs with AEI based SDG action descriptor.

Table 3.8: Pair-wise cross-view action recognition accuracy for the proposed approach on the IXMAS dataset

Train/Test View	**Cam0**		**Cam1**		**Cam2**		**Cam3**		**Cam4**	
	HMM	SVM	HMM	SVM	HMM	SVM	HMM	SVM	HMM	SVM
Cam0	-		88.1	**88.6**	**87.6**	86.5	93	94.7	86	**87.3**
Cam1	**89.6**	88.4	-		87	87.6	93.5	94.8	86.6	**87.8**
Cam2	88.4	**89.5**	**91.4**	90.5	-		93.6	94.1	85.8	**86.3**
Cam3	88.6	**90.2**	87	**88.3**	**88.9**	88	-		84.2	**86.5**
Cam4	87	**87.9**	86.5	**87.7**	86.1	**86.2**	91.9	92.8	-	
Avg.	88.4	**89**	88.25	**88.8**	**87.4**	87	93	**94.1**	85.65	**87**



(a)

(b)

(c)

(d)

Figure 3.19: Confusion matrix for four publicly available datasets (a) Weizmann (b) KTH (c) Ballet (d) IXMAS dataset (test view - cam2) using SVM classifier

Table 3.9: Comparison with other state-of-the-arts on Weizmann Dataset

Method	**Parameters**		
	Classifiers	***Test scheme***	***ARA (%)***
Contour Points [190]	KNN	LOSO	92.8
BoCP [191]	SVM	LOSO	97.78
BST [192]	SVM	LOPO	95.42
SC-STV [193]	NNC	-	96.3
CA [195]	CNN-HMM	-	89.2
VCHA [196]	NN	-	98.52
Proposed method	**HMM/SVM**	**LOO**	**96.2/97.5**

Table 3.10: Comparison with other state-of-the-arts on KTH Dataset

Method	**Parameters**		
	Classifiers	***Cross Validation scheme***	***ARA (%)***
AMI [197]	SVM	-	93.30
PDE [198]	KNN	LOO	92.6
BST [192]	SVM	LOPO	93.14
NSF [199]	KNN	LOO	94.49
Hankelets [200]	SVM	LOO	95.89
SC-STV [193]	NNC	LOO	94.33
SGF [194]	AdaBoost	LOO	95.17
CA [195]	CNN-HMM	-	94.43
VCHA [196]	NN	-	93.16
DTD [201]	CNN-LSTM	-	96.8
Proposed method	**SVM, HMM**	**LOO**	**96.6, 95.8**

Table 3.11: Comparison with other HAR state of-the-art methods on Ballet dataset

Method	**Parameters**		
	Classifiers	***Cross Validation scheme***	***ARA (%)***
MLMF [202]	Adaboost	*LOOCV*	*51*
SLTM [203]	S-CTM	*LOO*	*91.3*
Cuboid-LMPF [185]	RSR	*LOO*	*91.1*
Chaotic Invariants [204]	RVM	*LOO*	*90.8*
DBW [205]	SVM	*LOO*	*91.1*
HSC [58]	SVM-NN	*LOOCV*	*94.0*
BS-SDG [57]	SVM-NN	*LOOCV*	*94.5*
Proposed Method	**SVM, HMM**	***LOOCV***	***95.62, 94.3***

Table 3.12: Comparison with other HAR state-of-the-art methods on IXMAS dataset

Method	***Input***	***Actions***	***Views***	***Accuracy***
CDF-UBM [206]	*Images*	*12*	*4*	*88.2*
MHV [186]	*Silhouettes*	*11*	*5*	*93.3*
3DHOG [207]	*Images*	*11*	*5*	*83.5*
LKSS [208]	*Images*	*12*	*5*	*86.21*
Hankelets [200]	*Images*	*11*	*5*	*90.5*
KFSO [171]	*AEI*	*12*	*5*	*85.8*
STF [209]	*Images*	*11*	*5*	*86.9*
Proposed Method	***AEI***	***13***	***5***	***89.18***

It is noticed from Table 3.12, that the proposed framework performed superior than the other state-of-art methods [171] [209] [210] [208] with 89.18% recognition accuracy. Hankelets [200] obtained little higher recognition accuracy due to view-invariant property of Hanklets which do not carry Spatio-temporal information of an activity. The Hanklet approach finds ‘Pick Up’ activity hardest in IXMAs dataset to recognize with an average accuracy of 86.5% due to severe occlusion. However, the proposed Spatio-temporal approach can recognize ‘Pick up’ with 100% accuracy. Motion History Volume based action description [186] used Fourier-magnitudes and cylindrical coordinates, to represent translational and rotation invariant motion templates around the z-axis. However, a single template is not sufficient to represent all kinds of motion i.e. ‘turn-around’ can be misinterpreted as ‘single side step’ in small steps. Therefore, to remove this interclass similarity single motion template should be replaced with temporal networks of motion. The accuracy achieved by CDF-UBM [206], is very close to the proposed algorithm due to the use of alike spatiotemporal context distribution interest points. Though, the proposed algorithm obtained higher accuracy due to additional scale and view-invariant STIP used with AEI based SDG feature. Computational efficiency of the framework is analyszed in terms of time taken by each step for a single interation and given in Table 3.13. The computational complexity is an important parameter for the practical implementation of any method. Therefore, the approach should be efficient and simple to handle the computational complexity. We consider the proposed hybrid feature computations to illustrate the computational efficiency. For experiments, the proposed work is implemented using image processing toolbox, however for comparison with HPM+TM model [187], it is implemented using Matlab2017b with MatConvNet toolbox on a single NVIDIA GeForce 940M Graphics Card, 8GB RAM, Intel core i5 Processor. The time performance of the proposed work is evaluated by dividing the hybrid feature vector formation into five major steps: entropy based silhouette extraction, Average Energy Image (AEI) formation, SDG descriptor computation, DoG based STIPs extraction and Quantized Code-word generation. Execution time for each dataset for feature extraction is explained in Table 3.13, which confirms that the hybrid feature vector can be produced in an affordable time.

Table 3.13: Execution time of key steps for hybrid feature vector

Key Steps / Dataset	Entropy based silhouette extraction (per frame)	AEI formation (per video sequence)	SDG descriptor (per video)	DoG based STIPs detection (per frame)	Quantised Code word generation (Per Video)
Weizmann	0.690 sec	0.300 sec	0.137sec	2.43 sec	0.526 sec
KTH	0.473 sec	0.296 sec	0.137sec	1.58 sec	1.15 sec
Ballet	0.559 sec	0.344 sec	0.137sec	2.66 sec	0.44 sec
IXMAS	0.211 sec	0.121 sec	0.137sec	2.04 sec	1.934 sec
Average Time	0.483 sec	0.265 sec	0.137sec	2.177 sec	1.012sec

3.3 Significant Outcomes

The outcomes of this chapter are twofold. Firstly, this chapter addresses the problem of poor human action identification due to the illumination and scale variations in videos. Secondly, it targets the problem of elderly health care by developing a robust automated abnormal human action identification system using handcrafted model. The experimental results demonstrate some interesting observations, which are as follows:

- ***Dataset generation***- A new dataset, possessing five set of abnormal actions, is generated which is captured by Microsoft Kinect sensor v1. It consists of total 80 samples for five abnormal actions: ‘headache’, ‘chest pain’, ‘fainting’, ‘backward fall’ and ‘forward fall’. It is made publically available for other researchers in order to extend the research work in this direction.
- Encryption of structural appearance and temporal motion of human pose as Average Energy Silhouette Images (AESI) support low storage capacity and low computational cost.

- A robust translation, scale and rotation invariant novel action descriptor, developed by unifying the properties of both $\mathcal{R}$ -transform and Zernike moment, outperforms similar silhouette based action recognition methods, as provided in Table 3.3, 3.4, and 3.5.

- Global and local features based action representation is able to handle the illumination, scale and view variations discriminately, in videos. It is observed from the experimental results that fusion of DoG based STIPs with its SDG descriptor, potentially leads to the significant improvements for the KTH and Ballet, and IXMAS dataset, despite different environmental conditions, high intra-class variations in terms of speed of action, spatiotemporal scaling, clothing, and view variations etc.

- For human silhouette extraction, entropy based texture segmentation method works quite well in binary texture scenes. However, the performance of silhouette extraction may get effected for complex or multi-textured scenes.

- It is observed that the performance of any human action identification system greatly depends on how the action features are engineered. In this chapter handcrafted features extraction procedures are discussed and their performances are evaluated. However, complexity involved in designing handcrafted action descriptors for the actions acquired in challenging environmental conditions increases for efficient action recognition in videos. Therefore, now researchers are delving towards deep features to handle the practical challenges involved in action recognition in videos more effectively. Chapter 4 discusses about deep features based action recognition methods in videos.

This chapter is based on the following works:

- **C. Dhiman,** 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.
- D. K. Vishwakarma, **C. Dhiman**, "A unified model for human activity recognition using spatial distribution of gradients and difference of Gaussian kernel" ***The Visual Computers***, Vol. 35, no. 11, pp. 1595-1613, 2019.

CHAPTER 4

DEEP LEARNING MODELS

The objective of this chapter is to describe two human action identification deep models using the concept of transfer learning and part-wise feature engineering approach followed by late fusion. The key components of this chapter include the extensive study for highly discriminating deep features' engineering which include transfer learning based view invariant depth human pose description, motion dynamics encryption as Dynamic Images (DIs), Residual Inception Attention driven CNN network (RIAC-Net) based part wise human action representation and weighted late fusion. The proposed deep models are supported by experimental validation, results discussion and comparative analysis of results with the similar state-of-the-arts.

4.1 View-invariant Deep Human Action Recognition model using motion and Shape temporal dynamics

Recognition of human actions for unknown views is a challenging task. In this section, we propose a view-invariant deep action recognition framework is proposed, which is a novel integration of two important action cues: motion and shape temporal dynamics (STD). The motion stream encapsulates the motion content of action as RGB Dynamic Images (RGB-DIs) which are processed by the fine-tuned InceptionV3 model. The STD stream learns long-term view-invariant shape dynamics of action using human pose model (HPM) based view-invariant features mined from structural similarity index matrix (SSIM) based key depth human pose frames. To predict the score of the test sample, three types of late fusion (maximum, average and product) techniques are applied on individual stream scores. Cross subject and cross-view validation schemes are used to evaluate the proposed work. Our algorithm outperforms the existing state-of-the-arts significantly that is reported in terms of accuracy, Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC). The detailed description of developed approach, the experimental results and their discussion are provided in the subsequent sections.

4.1.1 Proposed Approach

Schematic block diagram of the deep view-invariant RGB-D action recognition framework is demonstrated in Fig. 4.1. The architecture is designed by learning both motion, view-invariant deep shape of the object over a period. The motion content of the action is encrypted as dynamic images (DI), and the concept of transfer learning is used to understand the action in RGB videos with the help of InceptionV3. Geometric details of the shape of the object during actions are extracted as view-invariant Human Pose Model (HPM) [36] features which are learned in a sequential manner using one Bi-LSTM, and one LSTM layer followed by dense, dropout and softmax layers. Two streams are combined using a late fusion concept to predict the action.

4.1.1.1 Depth-Human Pose Model (HPM) based action descriptor

Depth shape representation of human pose preserves the information about relative positions of the body parts. In the proposed work, fine details of depth human pose irrespective of the viewpoint are represented as view-invariant HPM features. HPM model [36] has similar architecture to the AlexNet [211]. It is trained with synthetically generated multi-viewpoint action data from 180 viewpoints which are generated by fitting human pose models to the CMU motion capture data [188]. It makes HPM view invariant. To preserve the temporal structure of action, HPM features are learned over LSTM sequential model.

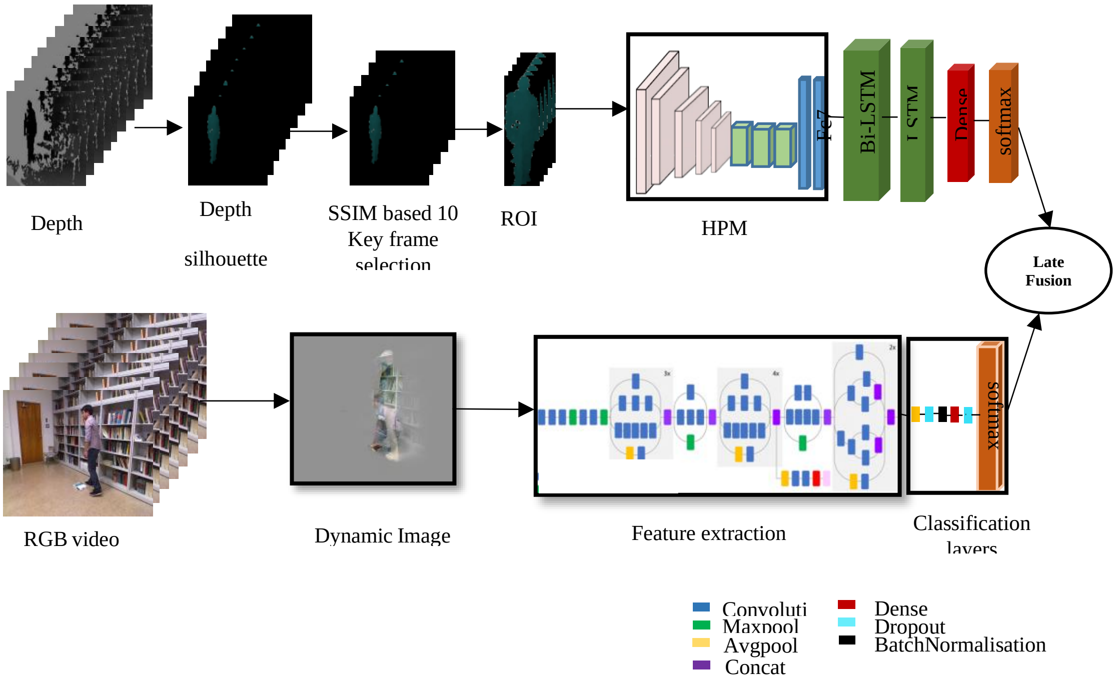


Figure 4.1: Schematic Block Diagram of the proposed approach

4.1.1.1.1 Self-Similarity Index Matrix (SSIM) based key frame Extraction

Initially a depth video with n no. of depth frames$\{f_1, f_2 \dots \dots, f_n\}$, is pre-processed by morphological operations, to obtain depth human silhouette to reduce the background noise. The redundant information in the video is removed by selecting key pose frames based on the Structural Similarity Index Matrix (SSIM) [212]. It computes the global structural similarity index value and local SSIM map for two consecutive depth frames. If there are small changes in a human pose with time during an action structural similarity index ($\mathbb{SSI}$) value is high. For distinct human poses, the $\mathbb{SSI}$ value is small. Mathematically SSIM value is defined as below:

$$\mathbb{SSI}(f_i, f_{i+1}) = [\mathcal{L}(f_i, f_{i+1})^{\alpha}] \times \left[\mathbb{C}(f_i, f_{i+1})^{\beta}\right] \times [\mathcal{S}(f_i, f_{i+1})^{\gamma}] \tag{4.1}$$

where $\mathcal{L}(f_i, f_{i+1}) = \frac{(2*\vartheta_{f_i}*\vartheta_{f_{i+1}}+\mathrm{K}_1)}{\vartheta_x^2+\vartheta_y^2+\mathrm{K}_1}$, $\mathbb{C}(f_i, f_{i+1}) = \frac{(2*\sigma_x*\sigma_y+\mathrm{K}_2)}{\sigma_x^2+\sigma_y^2+\mathrm{K}_2}$, $\mathcal{S}(f_i, f_{i+1}) = \frac{(\sigma_{xy}+\mathrm{K}_3)}{\sigma_x\sigma_y+\mathrm{K}_3}$

where $\vartheta_x, \vartheta_y, \sigma_x, \sigma_y, \sigma_{xy}$are the local means, variances and cross-variances for any two consecutive frames $f_i, f_{i+1}$and $\mathcal{L}(.)$,, $\mathbb{C}(.)$and $\mathcal{S}(.)$ are luminance, contrast and structural components of the pixels. Since depth images are not sensitive to luminance and contrast components, the exponents of $\mathcal{L}(.)$ and $\mathbb{C}(.) i.e. \alpha, \beta$ are set to 0.5 and exponent of structural component $\mathcal{S}(.)$, γ is set to 1. $\mathbb{SSI}$ value is computed for every two consecutive frames in a video and arranged in an ascending order with their respective frame numbers, in a vector Λ. First ten $\mathbb{SSI}$ values and corresponding frames numbers $i, i \in (1, n)$ are selected from the arranged vector Λ as key frames. The salient information of each selected key frames is extracted as region of interest (ROI) and resized to $[227 \times 227]$ images to transform into view-invariant HPM features composed as $fc7$ layer $[10 \times 4096]$ feature vector using HPM [36] model.

4.1.1.1.2 Model architecture and learning

In this paper shape temporal dynamics (STD) stream is designed to describe the long term shape dynamics of the action with deep convolutional neural network structure, whose architecture is similar to [36] except that we have connected the last $fc7$ layer with a combination of Bidirectional LSTM and LSTM layers. The architecture of our CNN follows:

$$Input(227,227) \rightarrow Conv(11,96,4) \rightarrow ReLU \rightarrow maxPool(3,2) \rightarrow$$
$$Norm \rightarrow Conv(5,256,1) \rightarrow ReLU \rightarrow maxPool(3,2) \rightarrow Norm \rightarrow Conv(3,256,1) \rightarrow$$
$$ReLU \rightarrow P(3,2) \rightarrow Fc6(4096) \rightarrow ReLU \rightarrow Dropout(0.5) \rightarrow Fc7(4096) \rightarrow$$
$$BiLstm(512) \rightarrow Lstm(128) \rightarrow ReLU \rightarrow Fc(.) \rightarrow softmax(.)$$

where $Conv(h, \mathscr{n}, \mathbb{s})$ is a convolution layer with $h \times h$ kernel size, $\mathscr{n}$ number of filters, $\mathbb{s}$ stride, $maxPool(h, \mathbb{s})$ is a max pooling layer of $h \times h$ kernel size and stride $\mathbb{s}$, $Norm$ is a normalization layer, $ReLU$ is a rectified linear unit, $Dropout(p)$ is Dropout layer with (p) dropout ratio, $Fc(\mathbb{N})$ is a fully connected layer with $\mathbb{N}$ no. of neurons. $BiLstm(\mathrm{O})$and $Lstm(\mathrm{O})$are Bidirectional Long short term memory(LSTM) layer, and one directional LSTM later respectively with 'O' output shape. Bidirectional LSTM layer is trained with weight regularizer 0.001 and recurrent dropout of 0.55 with the true return sequence for Bidirectional LSTM layer. Softmax layer is attached in the end of the network. Last fully connected layer is designed with 10, 30, and 60 neurons as output shape for NUCLA, UWA3D, and NTU RGB-D dataset respectively according to number of classes in the datasets.

The pre-trained HPM model is learned for view invariant synthetic action data for 399 types of human poses. Therefore, the proposed deep HPM based shape descriptor model is learned end to end with 80 epochs and 'Adam' optimizer. The $softmax$ layer will generate a probability vector $[1 \times n]$, where n is no. of classes, that shows the belongingness of the test sample to all the classes of the dataset.

4.1.1.2 RGB-Dynamic Image (DI) based action descriptor

In this section appearance and dynamics of a video is represented in terms of dynamic images (DIs), which are later used to learn pre-trained inceptionV3 architecture according to the dynamics of the action sequence. DIs focus mainly on the salient objects and motion of the salient object by averaging away the background pixels and their motion patterns, by preserving long-term action dynamics. In comparison to other sequence invariant temporal pooling strategies [213] [35], ARP emphasize the order (τ) of frame occurrence to extract complex long term dynamics of an action.

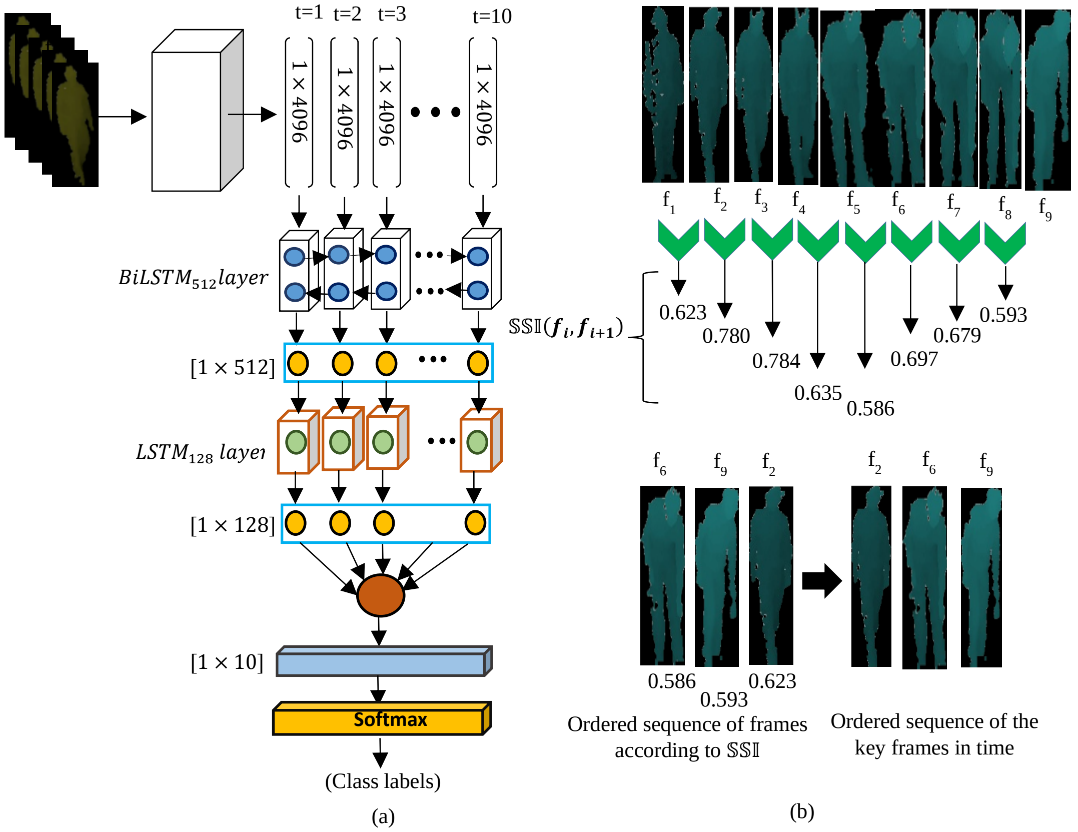


Figure 4.2: (a) Shape Temporal Dynamics (STD) stream design (b) SSIM based key feature extraction procedure is demonstrated for only nine frames as a test case considering α=0.5,β=0.5, γ=1

Construction of dynamic image depends on the ranking function that rank each frame in time axis. According to Fernando et al. [35], a video, i.e. $\{I_1, I_2, \dots\dots, I_N\}$ is represented as a ranking function$\varphi(I_t), t \in [1, N]$ where, $\varphi(.)$ function assigns a score $\mathscr{s}$ to each frame $I_t$at instance t to reflect the rank of each frame. The time average of $\varphi(I_t)$ up to time t is computed as $Q_t = \frac{1}{t}\sum_{i=1}^{t}\varphi(I_i)$ and$\mathscr{s}(t|\boldsymbol{r}) =< \boldsymbol{r},\ Q_t >$, where $\boldsymbol{r} \in R^r$ is a vector of parameters. Score for each frame is computed in such a manner that$\mathscr{s}(t_2|\boldsymbol{r}) > \mathscr{s}(t_1|\boldsymbol{r}), t_2 > t_1$. For which vector $\boldsymbol{r}$ is learned as a convex optimization problem using RankSVM [214]. The optimising equation is given as Eq. (4.2):

$$r^* = \partial(I_1, I_2, \ldots\ldots, I_N : \varphi) = argmin_r E(r), \tag{4.2}$$

where $E(r) = \frac{\lambda}{2}\|r\|^2 + \frac{2}{N(N-1)} \times \sum_{t_1 > t_2} \max\{0,1 - \mathscr{s}(t_2|\boldsymbol{r}) + \mathscr{s}(t_1|\boldsymbol{r})\}$ and $\partial(I_1, I_2, \ldots\ldots, I_N : \varphi)$ maps a sequence of N number of frames to r^*, also termed as rank pooling function, that holds the information to rank all the frames in the video. The first term in objective function $E(r)$ is the quadratic regularised used in support vector machines. The second term is a hinge-loss that counts the number of pairs $t_2 > t_1$ are falsely ranked by the scoring function $\mathscr{s}(.)$., if scores are not separated by at least unit margin i.e. $\mathscr{s}(t_2|\boldsymbol{r}) > \mathscr{s}(t_1|\boldsymbol{r}) + 1$. In the proposed work, learning of the ranking function for dynamic images construction is accelerated by applying approximate rank

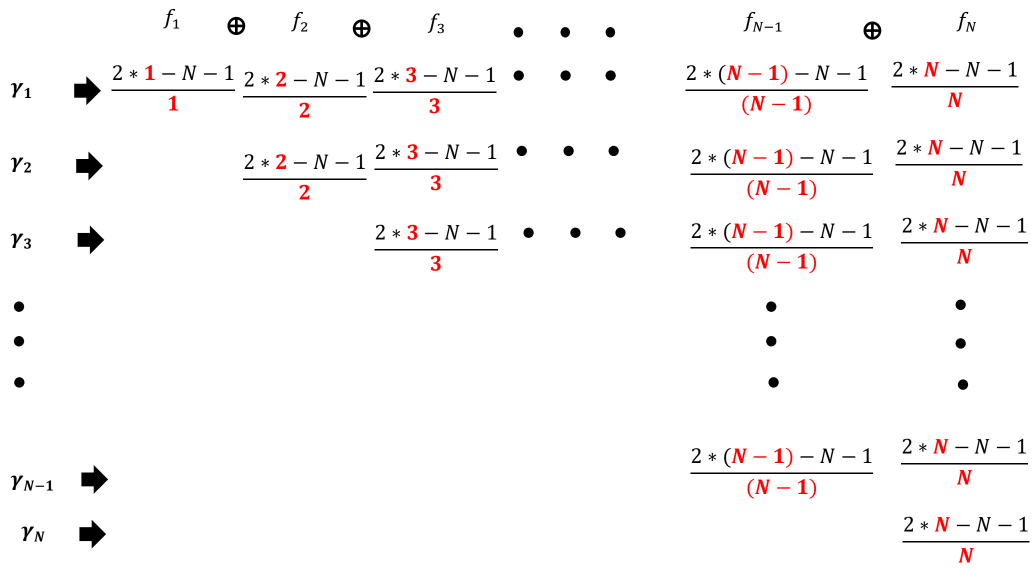

f_1 ⊕ f_2 ⊕ f_3 • • • f_{N-1} ⊕ f_N
γ_1
$\frac{2*\mathbf{1}-N-1}{\mathbf{1}}$
$\frac{2*\mathbf{2}-N-1}{\mathbf{2}}$
$\frac{2*\mathbf{3}-N-1}{\mathbf{3}}$
$\frac{2*(\mathbf{N-1})-N-1}{(\mathbf{N-1})}$
$\frac{2*\mathbf{N}-N-1}{\mathbf{N}}$
γ_2
γ_3
γ_{N-1}
γ_N

Figure 4.3: computation of γ_t parameter for fixed video length N; numbers in red show the dependency of γ_t on consecutive video frames $\in (\boldsymbol{i}, \boldsymbol{N})$

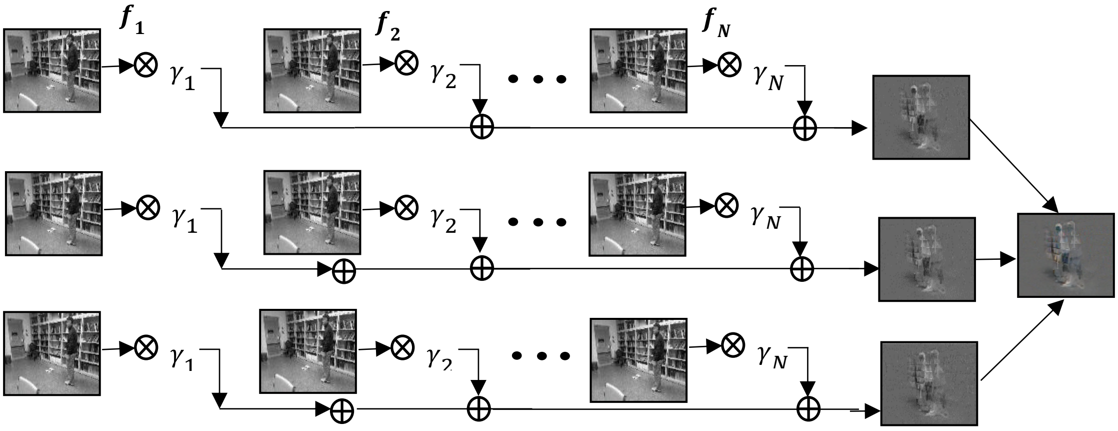


Figure 4.4: Dynamic image formation using Approximate Rank Pooling (ARP) [48]. First row: R-channel, Second row: G-channel, Third row: B-channel of each RGB video frame

pooling (ARP) [215]. It involves simple linear operations at pixel level, over the frames to rank them, which is extremely efficient and simple for fast computation. ARP approximates the rank pooling procedure by using gradient-based optimization in Eq. (4.3) as follow:

For $\boldsymbol{r} = \vec{0}, \boldsymbol{r}^* = \vec{0} - \eta \nabla E(\boldsymbol{r})|_{\boldsymbol{r}=\vec{0}}$ for any $\eta > 0$, (4.4)

where $\nabla E(r) \propto \sum_{t_2>t_1} \nabla \max\{0,1 - s(t_2|\boldsymbol{r}) + s(t_1|\boldsymbol{r})\}\,|_{d=\vec{0}}$

$$\nabla E(r) = \sum_{t_2>t_1} \nabla < \boldsymbol{r}, Q_{t1} - Q_{t2} > = \sum_{t_2>t_1} Q_{t1} - Q_{t2} \tag{4.5}$$

$$\boldsymbol{r}^* \propto \sum_{t_2>t_1} \left[\frac{1}{t_2}\sum_{i=1}^{t_2} \varphi_i - \frac{1}{t_1}\sum_{j=1}^{t_1} \varphi_j\right] = \sum_{t=1}^{T} \gamma_t\, \varphi_t \tag{4.6}$$

where $\gamma_t = 2(T - t + 1) - (T + 1)(h_t - h_{t-1})$, and $h_t = \sum_{i=1}^{t} \frac{1}{t}$ is the t^{th}harmonic number, $h_0 = 0$. Hence, rank-pooling function is re-written as:

$$\hat{\partial}(I_1, I_2, \dots\dots, I_N \colon \varphi) = \sum_{t=1}^{T} \gamma_t \varphi(I_t) \tag{4.7}$$

Therefore, ARP can be defined as a weighted sum of sequential video frames. The weights $\gamma_t, t\epsilon[1, N]$are pre computed for a fixed length video, using Eq. (4.7) as shown in Fig. 4.3. While computing γ_n, the order of occurrence of all the frames, for time $t \geq n$, are considered by computing a weight for each frame $\frac{2*i-N-1}{i}$, where $i \in [n, N]$. The computed weight value for each considered frame is summed to obtain single value of γ_n. Therefore, rank-pooling function can be directly defined by using individual frame features $\varphi(I_t)$ and $\gamma_t = 2(T - t + 1)$ as a linear function of time$t$, instead of computing the intermediate average feature vectors Q_t per frame to assign the score to rank the frames. The procedure of Approximate Rank Pooling (ARP) is shown in Fig. 4.4. Where each video frame is multiplied with the corresponding computed, $\gamma_t$weight i.e. $f_1$is multiplied with $\gamma_1$for every channel separately. R, G, and B channels of the dynamic image is obtained as weighted sum of R, G, and B -channels of each video frame respectively.

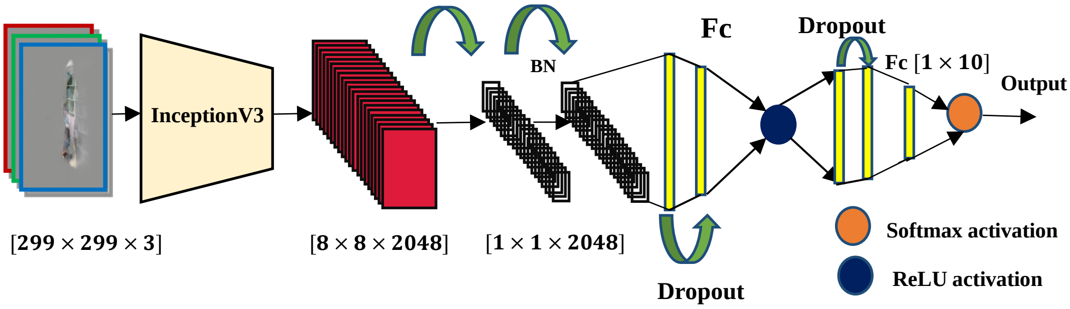


Figure 4.5: Layer Structure of the Motion Stream, GAP: Global Average Pooling, BN: Batch normalization

The size of the DI, so obtained, is same as original frame. To compute view invariant motion features of the action, the constructed DIs are passed through the motion stream as shown in Fig. 4.5, which is a combination of InceptionV3 architecture convolution layers followed by set of classification layers i.e. $Global\ Average\ Pooling\ Layer2D(\ ),\ BatchNormalisation(\ ), dropout(0.3),$ $dense(512,' Relu'), dropout(0.5), and\ Dense(10,'softmax')$ layers.Convolution features with vector shape of $8 \times 8 \times 2048$ is received as high dimensional representation of input image using pre-trained InceptionV3 model. Batch normalisation layer is used to maintain the internal covariate shift of hidden units' values to be minimal after 'ReLU' activations in $dense(512,'Relu')$ layer. Combination of Batch normalisation and Dropout layer helped to handle the overfitting phenomena without minimal loss of dropouts rather than only depending on dropout layer resulting in larger loss of weights. The layers of the motion stream are trained end to end for multiview datasets to update the weights of the InceptionV3 convoultion layers according to training samples. The best trained model weights so obtained for the highest achieved validation accuracy are used for testing of the sample to achieve the high recognition rate irrespective of view variations.

4.1.2 Experimental Work and Results

The performance of the proposed view-invariant human action recognition framework is tested on three publically available NUCLA multi-view action 3D dataset, UWA3D Depth dataset and NTU RGB-D activity datasets are used.

4.1.1.3 NUCLA multi-view action 3D Dataset

The Northern-UCLA multi-view RGB-D dataset [216] is captured by Jiang Wang and Xiaohan Nie in UCLA simultaneously from three different viewpoints using Kinect v1. The dataset covers 10 action categories performed by 10 subjects (l) pick up with one hand, (2) pick up with two hands, (3) drop trash, (4) walk around, (5) sit down, (6) stand up, (7) donning, (8) doffing, (9) throw, (l0) carry. Many actions share the same "walking" pattern before and after the actual action is performed, which increases the challenges offered by the dataset. To handle this inter-class similarity SSIM based ten depth key frames are selected and processed as 3D-HPM shape features. Some actions such as "pick up with on hand" and "pick up with two hands" are difficult to discriminate from different views. For cross view validation one view is used for testing and rest for training. For cross-subject validation, test samples are selected irrespective of viewpoint. The action samples of the dataset are given in Fig. 4.6(a).

4.1.1.4 UWA3D Multi view Activity-II Dataset

UWA3D multi-view activity-II dataset [217] is a large dataset which covers 30 human actions performed by ten subjects and recorded from 4 different viewpoints at different times using the Kinect v1 sensor. The 30 actions are: one hand waving, one hand punching, two hands waving, two hands punching, sitting down, standing up, vibrating, falling down, holding chest, holding head, holding back, walking, irregular walking, lying down, turning around, drinking, phone answering, bending, jumping jack, running, picking up, putting down, kicking, jumping, dancing, moping floor, sneezing, sitting down, squatting, and coughing. The four viewpoints are: (i) front, (ii) left, (iii) right, (iv) top. There exist small intra class similarity since large number of action classes are not recorded simultaneously. Sample images of UWA3D dataset from four different viewpoints are shown in Fig. 4.6(b).

4.1.1.5 NTU RGB-D Human Activity Dataset

NTU RGB+D action recognition dataset [218] is a largest and most complex cross-view RGB-D dataset for human activity analysis captured by 3 Microsoft Kinect v.2 cameras placed at three different angles: $-45^0, 0^0, 45^0$, simultaneously. It consists of 56,880 action samples including RGB videos, depth map sequences, 3D skeletal data, and infrared videos for each sample. The dataset consists of 60 types of actions performed by 40 subjects repeated two times. The sample frames of dataset are shown in Fig. 4.6(c). The resolution of RGB videos and depth maps is 1920×1080 and 512×424 respectively. We follow the standard cross subject and cross view evaluation protocol in the experiments, as specified in [218]. Under cross subject evaluation protocol, out of 40 subjects 20 subjects are selected for training and 20 subjects for testing. Under cross subject evaluation protocol, out of 40, 20 subjects are selected for training and 20 subjects for testing. Under cross view evaluation protocol, view 2 and view 3 are used as training views and view 1 is used as test view.

In the experiments, both motion stream and STD stream of the proposed deep framework are pre-trained end-to-end independently. For testing phase, the best-trained model is selected based on highest validation accuracy achieved. Under cross view validation scheme one view is used as test view and rest all views are used for training. In the training phase, the training samples are split in training samples and validation samples using 80-20 splitting strategy and Adaptive Moment Estimation (Adam) optimizer is used with (epochs, batch size, learning rate of the Adam optimizer) as $(80, 10, 0.0002)$. In the testing phase, the scores obtained from each stream for each test sample are fused using three late fusion mechanisms: maximum, average and product. The obtained performance of our approaches for cross subject and cross view validation scheme for NUCLA multi-view dataset and UWA3D II activity dataset is provided in Table 4.1, 4.2 and 4.3, which highlight the obtained highest accuracy of the proposed framework for each dataset. Where results are described in terms of motion stream, STD stream and proposed hybrid approach which stands for [DI_InceptionV3], [HPM_LSTM] and [HPM_LSTM + DI_InceptionV3] respectively. It is observed that for cross view and cross-subject validation both, late fusion scheme has produced remarkable results. The small variation in the accuracy for different views exhibits the view invariance property of the proposed framework. Whereas, in other state-of-the-arts [36] [219] [166], accuracy of the presented frameworks vary from view to view in the range of 10% that shows these state-of-the-arts are sensitive to different views. It is noticed that the novel integration of motion stream and STD stream of the proposed method has outperformed the recent works HPM_TM [36], HPM_TM+DT [135], NKTM [37]. Interestingly, our method achieves 91.3% and 83.6% average recognition accuracy which is about 9% and 10.86% higher than the nearest competitor HPM_TM+DT [135] when view 1 is considered for test view for both UWA3D II

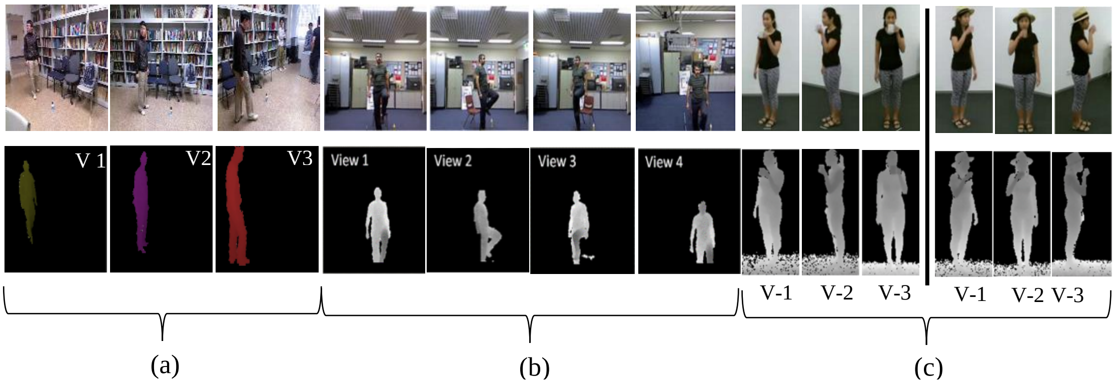


Figure 4.6: Sample Images of (a) NUCLA multi-view 3D Action Dataset (b) UWA3D II Multi-View Action Dataset and (c) NTU RGB-D Human Activity Dataset. Left group and right group of images in the Fig. 4.6(c) are recorded when the subject face camera sensor C-3, and C-2 respectively

activity dataset and NUCLA dataset. However, the obtained classification accuracy for NTU RGB-D dataset is not as good as obtained from other datasets due to the large variety of number of samples and their complexity in NTU RGB + D dataset. Viewpoint and large intra-class variations make this dataset very challenging. The performance of other work [170], Table 4.6, is comparatively better than the proposed framework for NTU RGB-D Activity dataset. It utilized the skeleton joints based action features to make prediction. However, the novel integration of motion stream and STD stream using late fusion has boosted recognition accuracy for all three multi-view datasets verified as ROC curves and AUC in Fig. 4.7 for individual test view of each multi-view dataset. From where it can be easily visualised that the hybrid approach based ROC curves are showing superior performance than the individual motion stream and STD stream based classification results which supports the fact that the fusion of the scores of two streams has resulted in increase in correct selection of true samples thereby improved true positive rate (TPR). At the same time, AUC values of the ROC curves help to understand and compare the ROC curves in a clearer way when they cross each other or nearly close to each other.

Table 4.1: **Cross Subject validation results in terms of ARA(%) for NUCLA and UWA3D II and NTU RGB-D Activity Dataset**

Dataset / Method		NUCLA dataset	UWA3D II dataset	NTU RGB-D dataset
Motion stream		93	82.6	62
STD stream		76	73.5	68.3
Proposed Hybrid Approach	Max	83	81.8	71.6
	Avg	84.5	79.6	75.7
	Mul.	**87.3**	**85.2**	**79.4**

Table 4.2: Cross View validation results in terms of ARA for NUCLA Multi-View Action 3D Dataset

Training/ Test View		**[1,2]/3**	**[1,3] / 2**	**[2,3]/ 1**	**ARA**
Motion stream		86.29	76.42	70.6	77.77
STD stream		58.88	73.67	63.83	65.46
Hybrid	Max	91.73	85.43	79.72	85.68
	Avg	90.46	80.65	74.50	81.87
	Mul.	**93.12**	**89.94**	**85.36**	**89.47**

Table 4.3: Cross View validation results in terms of ARA for UWA3D Multi View Activity-II Dataset (%)

Training View	**[v1,v2]**		**[v1, v3]**		**[v1,v4]**		**[v2,v3]**		**[v2,v4]**		**[v3,v4]**		**mean**
Test View	**v3**	**v4**	**v2**	**v4**	**v2**	**v3**	**v1**	**v4**	**v1**	**v3**	**v2**	**v4**	
Motion Stream	87.4	81.2	78.1	85.5	73.9	79.4	82.6	73.1	81.6	72.4	83.5	81.1	79.98
STD stream	62.1	73.5	69.6	79.6	65.4	75.9	64.3	69.5	66.3	69.8	78.6	68.8	70.2
Hybrid Approach (max, avg, mul)	86.6	**85.3**	81.8	86.5	78.3	82.8	85.1	83.6	85.1	81.2	85.3	82.3	83.65
	73.2	78.8	75.4	81.3	79.9	81.4	79.4	77.3	79.4	80.9	84.1	**84.2**	79.6
	88.2	84.3	**82.6**	**88.6**	**80.5**	**83.2**	**88.9**	**84.6**	**93.9**	**85.2**	**91.2**	83.0	**86.18**

Computation time: The proposed view-invariant deep model outperformed the recent state-of-the-arts on multi-view NUCLA, UWA3D II and NTU RGB-D Activity Dataset by fusing motion stream and view-invariant Shape Temporal Dynamics (STD) stream information. Therefore, the proposed two stream deep architecture not only perform proficiently but also time efficient in comparison with the existing view invaraint deep recognition models. The experiments are performed on 24GB RAM, NVIDIA Geforce RTX 2080 Ti GPU. It does not demand computationally expensive training and testing phases, as shown in Table 4.7. The major reason behind small computation cost involved in training and testing phase, is the compact and competent representation of action. In motion stream, the entire video sequence is represented by a single DI and STD stream process the key human pose depth frame instead of all the frames in the action sequence.

Table 4.4: Comparison with other-state-of-the-arts on NUCLA Multi-View Action 3D Dataset

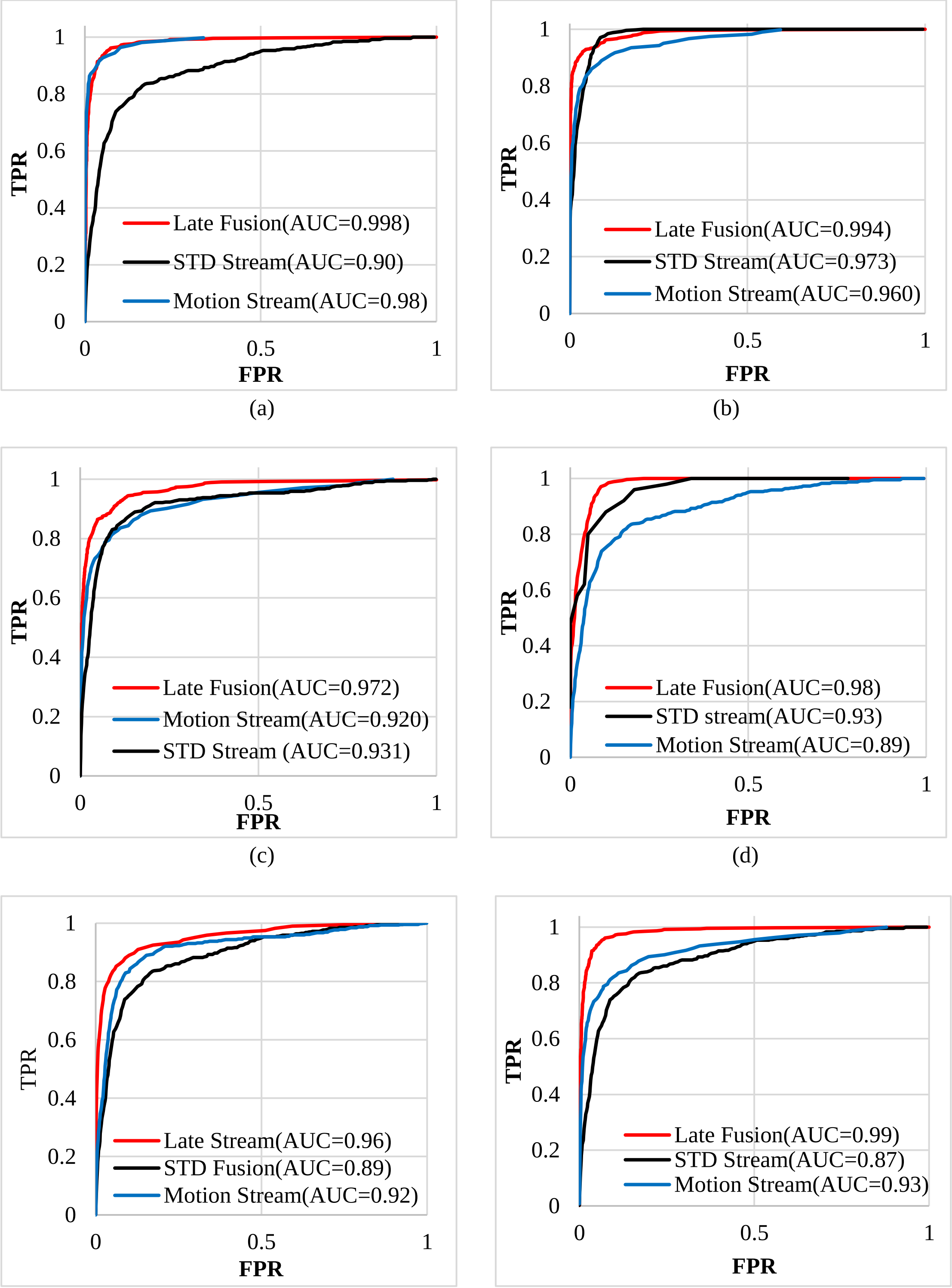
Train-Test View Methods	Data Type	[1,2]/ 3	[1,3] /2	[2,3] /1	Mean
CVP [220]	RGB	60.6	55.8	39.5	52
nCTE [221]	RGB	68.6	68.3	52.1	63
NKTM [37]	RGB	75.8	73.3	59.1	69.4
HOPC+STK [217]	Depth	80	-	-	-
HPM_TM [36]	Depth	92.2	78.5	68.5	79.7
HPM_TM+DT [135]	RGBD	92.9	82.8	72.5	82.7
HPM [36]	Depth	85.21	78.57	71.96	78.58
Motion Stream	RGB	86.29	79.7	70.6	77.77
STD stream	Depth	89.96	81.37	75.12	82.15
Proposed Hybrid Approach	RGBD	**93.12**	**89.94**	**85.36**	**89.47**

Table 4.5: Comparison with other state-of-the-arts in terms of ARA (%) on UWA3D Multi-View Activity-II Dataset

Train-Test Methods	Data Type	[v1,v2]		[v1, v3]		[v1,v4]		[v2,v3]		[v2,v4]		[v3,v4]		ARA
		v3	v4	v2	v4	v2	v3	v1	v4	v1	v3	v1	v2	
DT [222]	RGB	57.1	59.9	60.6	54.1	61.2	60.8	71	59.5	68.4	51.1	69.5	51.5	60.4
C3D [219]	RGB	59.5	59.6	56.6	64	59.5	60.8	71.7	60	69.5	53.5	67.1	50.4	61
nCTE [221]	RGB	55.6	60.6	56.7	62.5	61.9	60.4	69.9	56.1	70.3	54.9	71.7	54.1	61.2
NKTM [169]	RGB	60.1	61.3	57.1	65.1	61.6	66.8	70.6	59.5	73.2	59.3	72.5	54.5	63.5
R-NKTM [37]	RGB	64.9	67.7	61.2	68.4	64.9	70.1	73.6	66.5	73.6	60.8	75.5	61.2	67.4
HPM$_{\text{(RGB+D)_}}$Traj [166]	RGBD	85.8	89.9	79.3	85.4	74.4	78	83.3	73	91.1	82.1	90.3	80.5	82.8
HPM_TM+DT [135]	RGBD	86.9	89.8	81.9	89.5	76.7	83.6	83.6	79	89.6	82.1	89.2	83.8	84.6
HPM [36]	Depth	58.61	69.92	65.42	73.11	61.33	69.98	61.02	63.18	61.19	61.47	73.99	64.31	65.29
Motion Stream	RGB	87.4	81.2	78.1	85.5	73.9	79.4	82.6	73.1	81.6	72.4	83.5	81.1	79.98
STD stream	Depth	62.1	73.5	69.6	79.6	65.4	75.9	64.3	69.5	66.3	69.8	78.6	68.8	70.2
Proposed Hybrid Approach	RGBD	86.6	**85.3**	81.8	86.5	78.3	82.8	85.1	83.6	85.1	81.2	85.3	82.3	83.65
		73.2	78.8	75.4	81.3	79.9	81.4	79.4	77.3	79.4	80.9	84.1	**84.2**	79.6
		88.2	84.3	**82.6**	**88.6**	**80.5**	**83.2**	**88.9**	**84.6**	**93.9**	**85.2**	**91.2**	83.0	**86.18**

Table 4.6: Comparison of with other state-of-the-arts in terms of ARA (%) on NTU RGB-D Activity Dataset

Method	Data type	Cross subject validation	Cross view validation
Skepxel$_{\text{loc+vel}}$ [170]	Joints	**81.3**	**89.2**
STA-LSTM [223]	Joints	73.4	81.2
ST-LSTM [224]	Joints	69.2	77.7
HPM$_{\text{(RGB+D)_}}$Traj [166]	RGB-D	**80.9**	**86.1**
HPM_TM+DT [135]	RGB-D	77.5	84.5
Re-TCN [225]	Joints	74.3	83.1
dyadic [226]	RGB-D	62.1	-
DeepResnet-56 [227]	Joints	78.2	85.6
HPM [36]	Depth	65.8	70.9
Motion Stream	RGB	62	68.7
STD stream	Depth Maps	68.3	72.4
Proposed Hybrid Approach	RGB-D (max fusion)	71.6	79.8
	RGB-D (late fusion)	75.7	83
	RGB-D (product fusion)	**79.4**	**84.1**


TPR
FPR
1
0.8
0.6
0.4
0.2
0
0
0.5
1
Late Fusion(AUC=0.998)
STD Stream(AUC=0.90)
Motion Stream(AUC=0.98)
Late Fusion(AUC=0.994)
STD Stream(AUC=0.973)
Motion Stream(AUC=0.960)
Late Fusion(AUC=0.972)
Motion Stream(AUC=0.920)
STD Stream (AUC=0.931)
Late Fusion(AUC=0.98)
STD stream(AUC=0.93)
Motion Stream(AUC=0.89)
Late Stream(AUC=0.96)
STD Fusion(AUC=0.89)
Motion Stream(AUC=0.92)
Late Fusion(AUC=0.99)
STD Stream(AUC=0.87)
Motion Stream(AUC=0.93)

(a)

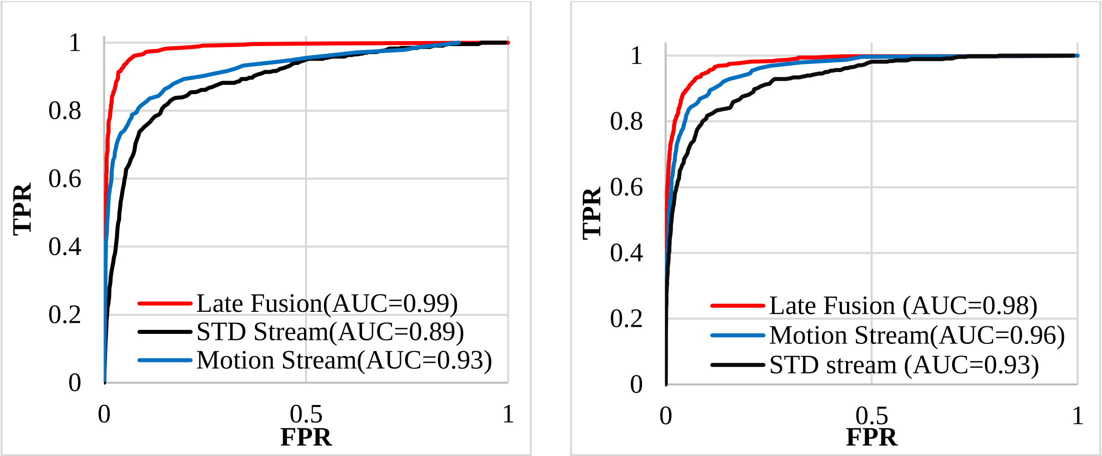
(b)

(c)

(d)

(e)

(f)



(g) (h)

Figure 4.7: Performance evaluation of the proposed framework for NUCLA multi-view dataset (a)-(c), UWA3D dataset (d)-(g) and NTU RGB-D Activity dataset (h) in terms of ROC curve and area under the curve (AUC)

Table 4.7: Average Computation Speed (frame per sec: fps)

Method	Training	Testing
NKTM [169]	12fps	16fps
HOPC [217]	0.04fps	0.5fps
HPM+TM [36]	22fps	25fps
Ours	**26fps**	**30 fps**

4.2 Part-wise Spatio-temporal Attention Driven CNN based 3D Human Action Recognition

There exist a wide range of intra-class variations of the same actions and inter-class similarity among the actions, at the same time, which makes the action recognition in videos very challenging. In this section, a novel skeleton-based part-wise Spatio-temporal CNN – RIAC Network-based 3D human action recognition framework is presented to visualize the action dynamics in part wise manner and utilize each part for action recognition by applying weighted late fusion mechanism. Part-wise skeleton-based motion dynamics helps to highlight local features of the skeleton which is performed by partitioning the complete skeleton in five parts- Head to Spine (HS), Left

Leg (LL), Right Leg (RL), Left Hand (LH), Right Hand (RH). The RIAC-Net architecture is greatly inspired by the InceptionV4 architecture which unified the ResNet and Inception based Spatio-temporal feature representation concept and achieving the highest top-1 accuracy till date. To extract and learn salient features for action recognition, attention driven residues are used which enhance the performance of residual components for effective 3D skeleton-based Spatio-temporal action representation. The robustness of the proposed framework is evaluated by performing extensive experiments on three challenging datasets such as UT Kinect Action 3D, Florence 3D action Dataset, and MSR Daily Action3D datasets, which consistently demonstrate the superiority of our method.

4.2.1 Proposed Approach

This section discusses about skeleton-based deep human action recognition framework. It describes the formation of Compact Action Skeleton Sequence formation and proposed RIAC-Net architecture design in detail which includes three main steps-formation of Spatial-Temporal Convolution Features (STCF), defining Attention Driven Residual Block (ADRB), and lastly learning part-wise RIAC-Net-based action features and to ensemble the predictions per part using weighted fusion scheme.

4.2.1.1 Compact Action Skeleton Sequence (CASS) generation

CASS is basically, a projection of each frame 3D coordinates of skeleton joints on the image frame which describes the spatial variation of the human skeleton pose during the action. The temporal details about the sequence of human poses are encrypted by using different colour coding for skeletons in such a way that colour of the skeletons changes with time to exhibit the sequence of occurrence of frames. To exploit the discriminative local features of the actions, the generated CASS are further divided into five significant parts: i) head to spine (HS) ii) left leg (LL) iii) right leg (RL) iv) left hand (LH) and v) right hand (RH). Therefore, CASS is defined for FS as well as for other five parts of the skeleton - {HS, RL, LL, RH, LH}.

Let an action video has n no. of frames $\{f_1, f_2 \dots \dots ., f_n\}$, and each frame possess a human skeleton with k no. of skeleton joints $i.e.$ $\{(J_{x1}, J_{y1}, J_{z1}), (J_{x2}, J_{y2}, J_{z2}), . \dots . (J_{xk}, J_{yk}, J_{zk})\}$, $k\epsilon[1,20]$. According to the configuration of skeleton joints in Figure 1, skeleton joints are partitioned in five groups as: $HS = \{J_4, J_3, J_2, J_1\}$, $LL = \{J_{13}, J_{14}, J_{15}, J_{16}\}$, $RL = \{J_{17}, J_{18}, J_{19}, J_{20}\}$, $LH = \{J_5, J_6, J_7, J_8\}$, $RH = \{J_9, J_{10}, J_{11}, J_{12}\}$. To generate CASS, the variations in joint coordinates of each group are sketched, mathematically defined as Eq. (4.5):

$$CASS_p = \begin{pmatrix} J_k^t & \cdots & J_k^n \\ \vdots & \ddots & \vdots \\ J_{k+4}^t & \cdots & J_{k+4}^n \end{pmatrix} \tag{4.5}$$

where k is the first skeleton coordinate number of each group, p is the partition label $\exists\ p\epsilon(HS, LL, RL, LH, RH)$, n is the number of frames in the action video. From the sample images of part-wise CASSs, for two actions- waving hand and sitting down, of the Florence3D Action dataset, in Fig. 4.8, it is observed that a different amount of motion is associated with each part of the skeleton for an action resulting in unique patterns for each action. The part wise feature extraction and learning highlight the local dynamics of the skeleton. However, if the complete skeleton is processed to extract spatial deep features, the prominent movements in the action would be subsided.

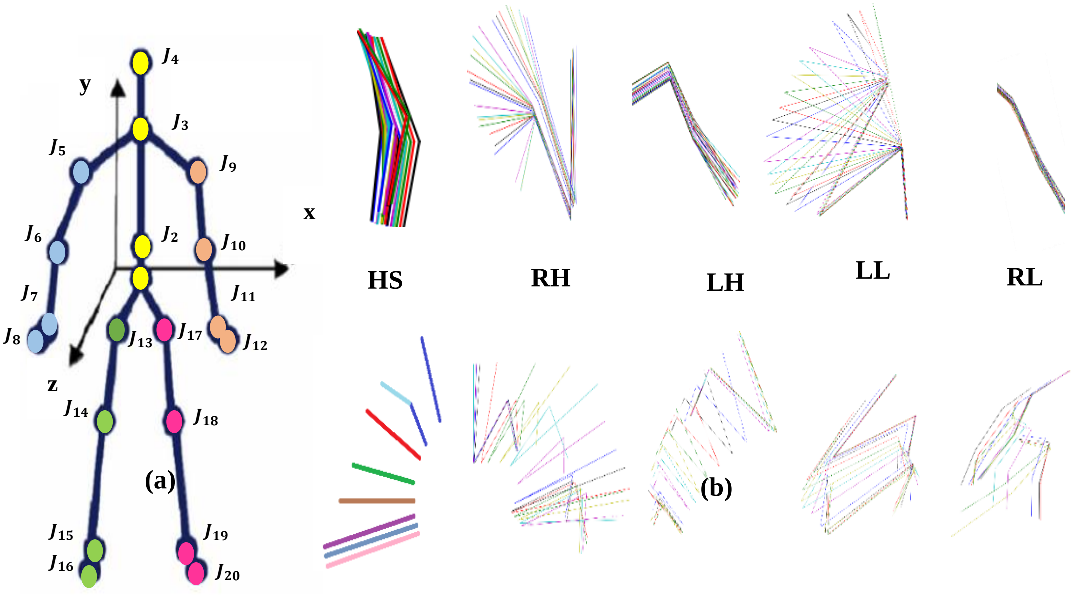


Figure 4.8: (a) Configuration of skeleton joints - Head to Spine (HS) joints: Yellow, LH joints: Blue, RH joints: Brown, LL joints: Green, RL joints: Pink (b) Part wise CASS formation for two actions – 'waving hand' and 'sitting down'

4.2.1.2 Skeleton based action recognition with RIAC-Net

The training of Inception networks with residual connections has accelerated significantly, resulting in outperforming the similarly expensive inception networks without residual connections [17]. Therefore, to solve the problem of skeleton-based action recognition for large inter-class similarity Residual Inception Attention-based Convolution Network (RIAC-Net), Fig. 4.9, is designed which is majorly divided into two parts- 'Spatial-Temporal Convolution Features' (STCF) and 'Attention Driven Residual Block' (ADRB).

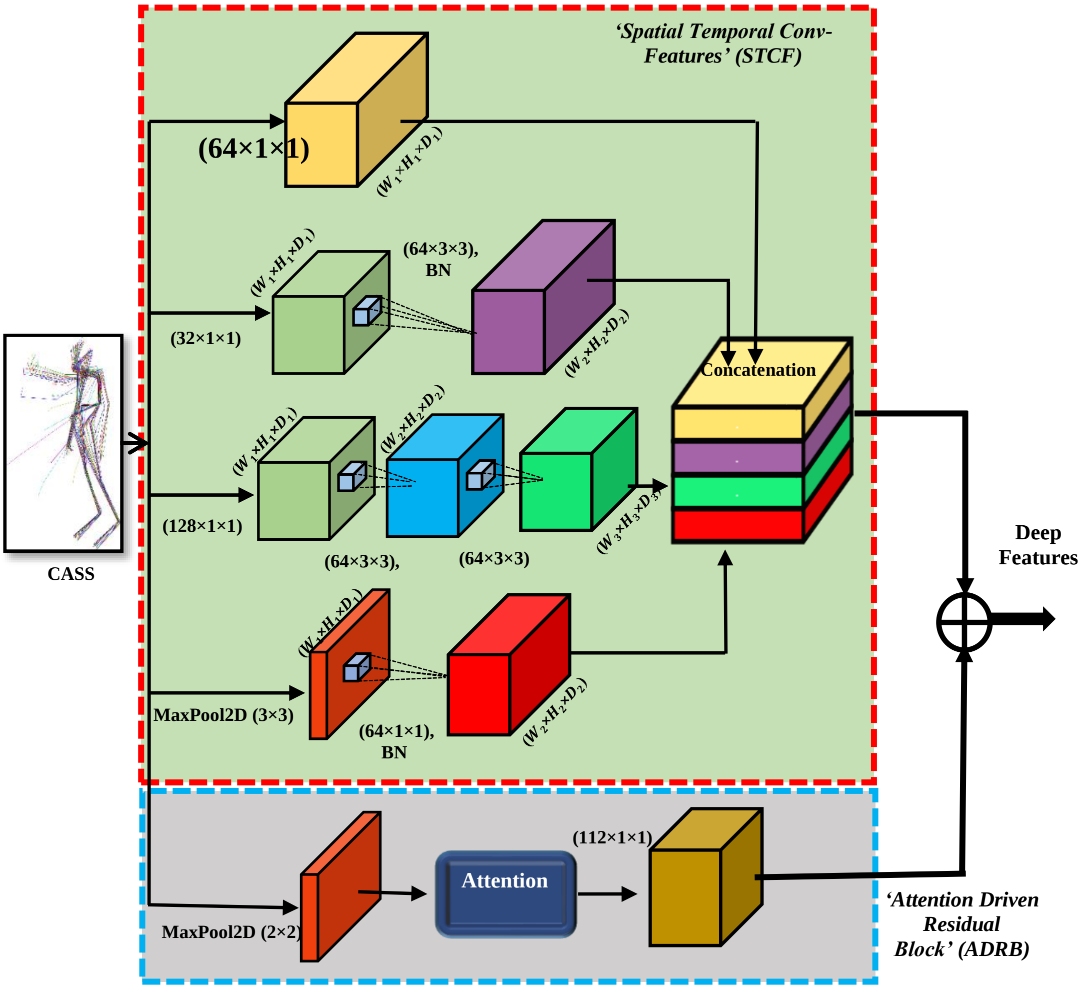

'Spatial Temporal Conv-Features' (STCF)
(64×1×1)
$(W_1 \times H_1 \times D_1)$
(64×3×3), BN
(32×1×1)
$(W_2 \times H_2 \times D_2)$
Concatenation
(128×1×1)
(64×3×3),
(64×3×3)
$(W_3 \times H_3 \times D_3)$
CASS
Deep Features
MaxPool2D (3×3)
(64×1×1), BN
Attention
(112×1×1)
MaxPool2D (2×2)
'Attention Driven Residual Block' (ADRB)

Figure 4.9: Proposed Residual Inception Attention-based Convolution Network (RIAC-Net) block diagram

4.2.1.2.1 Spatial-Temporal Convolution Features (STCF)

Salient features in an image, generally, can have an extremely large variation in size i.e. covering a major section of the image or small section. Convolution helps to recover Spatio-temporal features only with the right selection of kernel size. A large convolution kernel has a large receptive field that highlights the globally distributed information and a smaller convolution kernel is preferred for locally distributed information. Use of multiple sized kernels in convolution filters i.e. $(1 \times 1, 3 \times 3, 5 \times 5)$ is an efficient solution to select the appropriate kernel size for good convolution features [228].

Table 4.8: Description of RIAC-Net –STCF Block Architecture Parameters

RIAC-Net Branches	No. of filters	Kernel size/stride	Input size $(W_I \times H_I \times D_I)$	Output size $(W_O \times H_O \times D_O)$
Branch 1	Conv, 64	$(1 \times 1)/2$	$(224 \times 224 \times 3)$	$(112 \times 112 \times 64)$
Branch 2	Conv, 32	$(1 \times 1)/1$	$(224 \times 224 \times 64)$	$(224 \times 224 \times 32)$
	Conv, 64	$(3 \times 3)/2$	$(224 \times 224 \times 32)$	$(112 \times 112 \times 64)$
Branch 3	Conv, 128	$(1 \times 1)/1$	$(224 \times 224 \times 3)$	$(224 \times 224 \times 128)$
	Conv, 64	$(3 \times 3)/1$	$(224 \times 224 \times 128)$	$(224 \times 224 \times 64)$
	Conv, 64	$(3 \times 3)/2$	$(224 \times 224 \times 64)$	$(112 \times 112 \times 64)$
Branch 4	$Maxpool$	$(2 \times 2)/1$	$(224 \times 224 \times 3)$	$(112 \times 112 \times 3)$
	Conv, 64	$(1 \times 1)/1$	$(112 \times 112 \times 3)$	$(112 \times 112 \times 64)$

It essentially widens the network size and also computationally less expensive than deeper networks. It is the basic concept behind building the inception blocks [228] that targeted large size variations of spatial features. The convolution filters are made computationally more efficient by factorising (5×5) filters with two (3×3) filters in STCF. Description of the parameters of STCF block is provided in Table 4.8. It can be notified that equal-sized features $(W(112) \times H(112) \times D(64))$ are generated from all four branches of STCF block. And the final STCF feature vector is obtained by stacking STCF branch wise convolution features as $STCF_{FV} = \{STCFV_{FVi}\}$, where $i \epsilon [1,4]$ and each branch vector is constructed with dimension $(W_i \times H_i \times D_i) = (112 \times 112 \times 64)$ that results in $STCF_{FV}$ with $[448 \times 448 \times 64]$ dimesion.

4.2.1.2.2 Attention Driven Residual Block (ADRB)

The key structure of residual units allows input to the unit to propagate from first layer to the last layer of the network directly and gradients to propagate from the loss layer to any previous layer by skipping the midway weight layers during backpropagation, which helps to handle the vanishing gradient problem to a great extent. Hence, the idea of residuals [229] proliferated the performance of deep networks by adding the identity function. The effect can be further enhanced by adding salient features instead of the identity function directly. It is implemented by using Attention Driven Residual Block (ADRB), shown in Fig. 4.10 (a) and (b), where the attention block [230] tends to extract a spatial attention map by utilizing the inter-spatial relationship of features. The residual units are defined as follows:

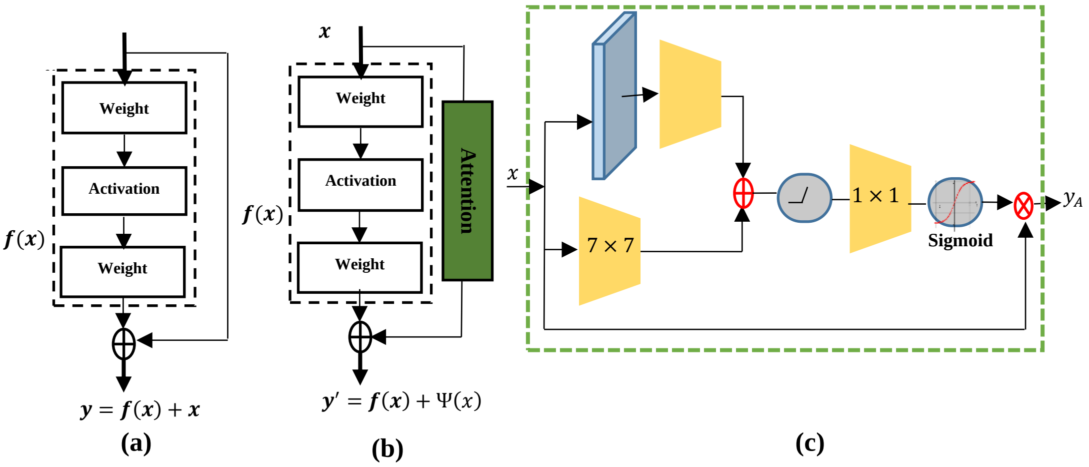


Figure 4.10: Illustration of Attention Driven Residual Block architecture (a) Basic Residual Block (b) Attention Driven Residual Block (c) Attention block

$$y = \sigma_1\big(x + \mathcal{F}(x; \mathcal{W})\big) \tag{4.6}$$

where x and y are the input and output of the RIAC-Net architecture, $\sigma_1(.) \equiv$ ReLU [231], and $\mathcal{F}$ is a non-linear residual mapping function for input x which is formulated as follows:

$$\mathcal{F}(x; \mathcal{W}) = STCF_{FV}(x; \mathcal{W}_k) \tag{4.7}$$

where is $\mathcal{W} = \{\mathcal{W}_k, 4 \leq k \leq 1\}$ for each convolution branch of STCF block. The residual unit is modified as

$$y' = \sigma_1\big(\Psi(x) + \mathcal{F}(x; \mathcal{W})\big) \tag{4.8}$$

And $\Psi(x) = x * \sigma_2\{f^{1\times1}\{\sigma_1[f^{7\times7}(x) + f^{1\times1}(\Lambda(x))]\}\}, \ \Psi\epsilon\mathbb{R}^{W\times H}$ (4.9)

where $\sigma_1$is a ReLU function and σ_2 is a sigmoid function, $f^{1\times1}(.), f^{7\times7}(.)$ are non-linear convolution layers with 1×1 and 7×7 kernel sizes and $\Lambda(.)$ is a 2D max-pool layer with 2×2 kernel size. The '*' and '+' are multiplicative [232] and additive [233] attention operator. To extract salient features additive attention method performs better for large dimensional input features [234] whereas the multiplicative attention method holds fast computations and also memory-efficient due to the matrix multiplication. Therefore, at the input stage additive attention operator is applied to handle larger input dimension than the later one, as shown in Fig. 4.10 (c).

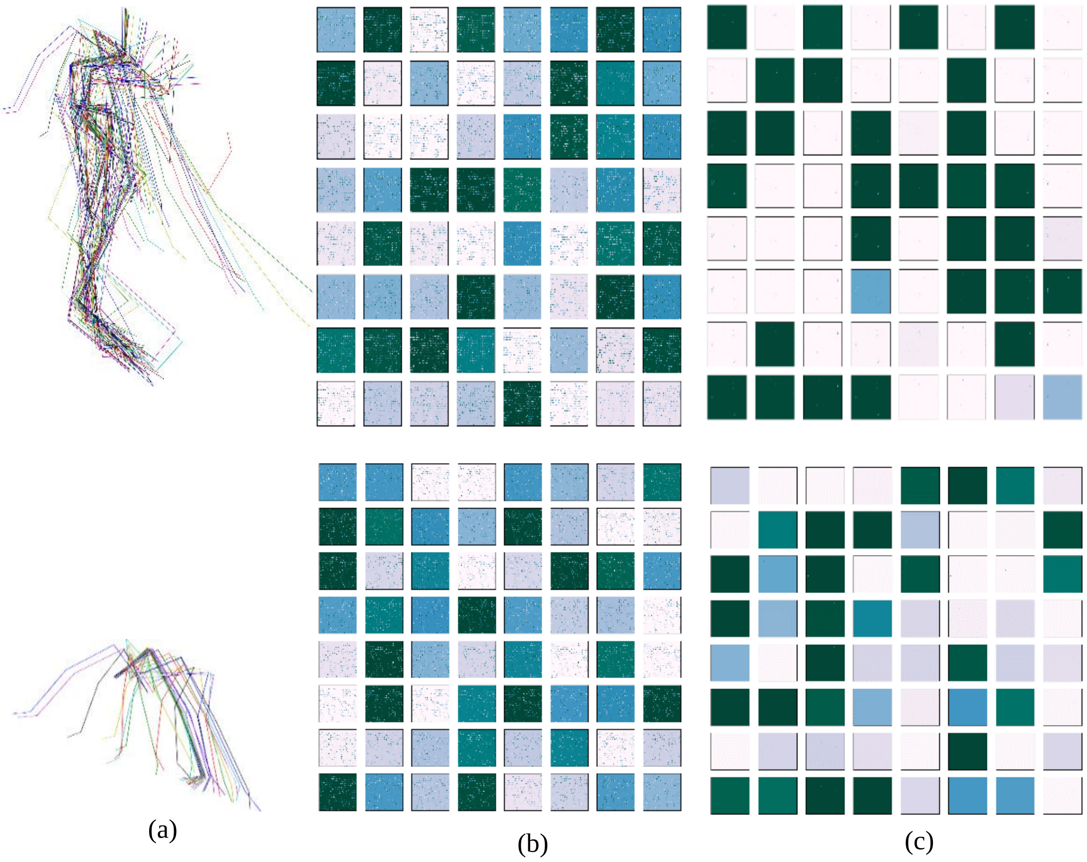


Figure 4.11: Illustration of Residue and Attention driven residue activation maps (8×8) for complete CASS of an action (a) whole Skelton of CASS, (b) and (e) residue activation maps, (c) and (f) attention driven residue activation map (d) left leg (LL) C

The two different sized kernels in the convolution layer $f^{\ k\times k}$ cover the features on the coarse spatial grid level and finer grid level which collectively helps to identify relevant features and disambiguate the task irrelevant features in x. Sample images of the residue and attention driven residue activation maps of the action, for complete CASS and part wise CASS, are shown in Fig. 4.11**.** It is clearly observed that the residue branch extracts noisy data whereas attention driven residue branch highlight the salient information about the action resulting in improved recognition performance.

4.2.1.2.3 Learning of part-wise RIAC-Net-based action descriptors

Late fusion approach works better than the early fusion scheme [235] at the cost of additional learning attempts. Therefore, RIAC-Net-based action descriptor is designed and learnt for each part-wise CASS, individually, using the combination of global average pooling(GAP), batch normalisation (BN), Long Short Term Memory (LSTM) layers, dropout layer with 0.2 dropout probability and dense layer. The final prediction is given by fusing the learnt part-wise CASS based predictions using weighted fusion scheme, as shown in Fig. 4.12. The best recognition performance for a specific weight combination i.e. $\{w_i\}\exists i \in [1,5]$, is reported finally. To learn the unique patterns of the part wise RIAC-Net based action features two consecutive LSTM layers are used in such a way that output gate $\boldsymbol{h_t^2}$of former LSTM layer is fed to the input gate $\boldsymbol{i_t^2}$of later LSTM layer. Let x_t^i , h_t^i , and C_t^i be input, output and cell state of the i^{th} LSTM layer, at instance t respectively. The sequence of flow of the signal from first LSTM layer to second LSTM layer is given by from Eq. (4.10) to (4.13) as follows:

$$h_t^1 = \sigma(W_0^1 * [h_{t-1}^1, x_t^1]+b_o) * \tanh(C_t^1) \tag{4.10}$$

$$C_t^1 = C_{t-1}^1 * f_t^1 + \widehat{C_t^1} * i_t^1 \tag{4.11}$$

$$x_t^2 = h_t^1 \; and \; h_t^2 = \sigma(W_0^2 * [h_{t-1}^2, x_t^2] + b_o) * \tanh(C_t^2) \tag{4.12}$$

$$C_t^2 = C_{t-1}^2 * f_t^2 + \widehat{C_t^2} * i_t^2 \tag{4.13}$$

where σ is the sigmoid operation, W_0^i is the weights of the output gate of the i^{th} LSTM layer, cell state $C_t^i$of the i^{th} layer is computed using both forget gate output as f_t^i and input gate output as i_t^t. $\widehat{C_t^i}$ control the amount of update required to current cell state C_t^i, according to the input x_t^i passed through input gate, i_t^i. The learnt vector h_t^2 with

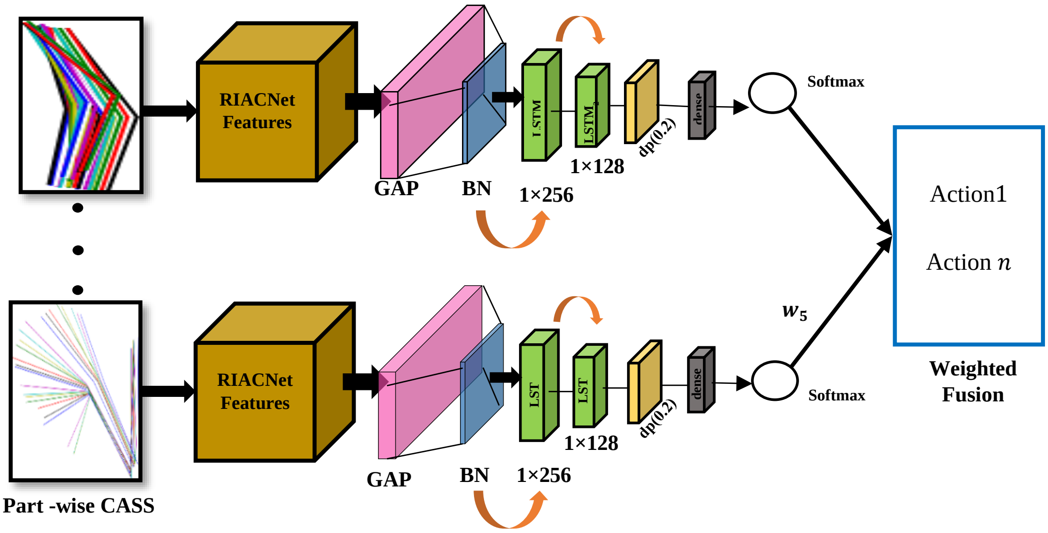


Figure 4.12: Description of proposed Part-wise Spatiotemporal and Attention Driven Residues based Learning for skeleton human action recognition

[1 × 128] dimension is fed to dropout layer to handle the overfitting problem followed by a dense layer and Softmax activation function. Predictions from each part-wise branch are fused using weighted fusion that utilises all possible combination of weights w_i to find the best set of weights, as shown follows:

$$P_c = \sum_{i=1}^{5} w_i \times p_i \, , c \in (1, n) \tag{4.14}$$

where n is the number of classes.

4.2.2 Experimental work and Results

The proposed framework is evaluated for three publically available 3D datasets- UT Kinect Action 3D and Florence 3D actions Dataset and MSR Daily Action3D datasets.

4.2.2.1 UT Kinect Action 3D dataset

UT Kinect Action 3D dataset [236] includes 10 human actions captured in indoor settings with single stationary Microsoft Kinect camera. The dataset includes an RGB image with 640×480 resolution, depth image with 320×240 and twenty 3D joints of a human skeleton per frame captured at 30 FPS. These actions are performed by 9 males and 1 female actor, each repeated two times, as shown in Fig. 4.13(a). Hence, it consists of a total of 200 (10×10×2) action samples with 6220 frames. The challenge lies in the fact that there exist viewpoint variations and high intra-class variations. The length of the sample actions ranges from 5 frames to 120 frames. Therefore, in the proposed work the number of frames for each action sample is made equal to 60 frames, before generating CASS representation of the action. It is performed by down-sampling the frames of the actions which possess more than 60 frames. And up-sampling is applied to the actions which possess less than 60 frames to maintain a symmetricity in the CASS generated using action frames. The sample RGB images are shown in Fig. 4.13(a) below. We use the skeleton representation of human actions for human action recognition.

4.2.2.2 Florence 3D Action dataset

Florence 3D Action dataset [237] dataset includes 9 actions performed by 10 actors as shown in Fig 4.13(b). It possesses total 215 skeleton action sequences captured by Microsoft Kinect SDK. The main challenges of this dataset are high inter class similarity and small intra class similarity among actions, the human object interaction.

4.2.2.3 MSR Action 3D Dataset

The MSR Action 3D dataset [238] consists of 20 actions, each performed by 10 subjects for three times. The dataset is divided into 3 subsets AS1, AS2, AS3 each include 8 actions, as shown in Table 4.9. It includes total 567 skeleton sequences. However, skeletons for 10 sequences are missing. Therefore, 557 action sequences are used for experimentation. Cross-subject evaluation protocol is used for experimentation. According to which, half of the dataset with 1, 3, 5, 7 and 9 subject ids, is used for training and another half of the dataset with 2, 4, 6, 8, 10 subject ids is used for testing for each action set-AS1, AS2, AS3.

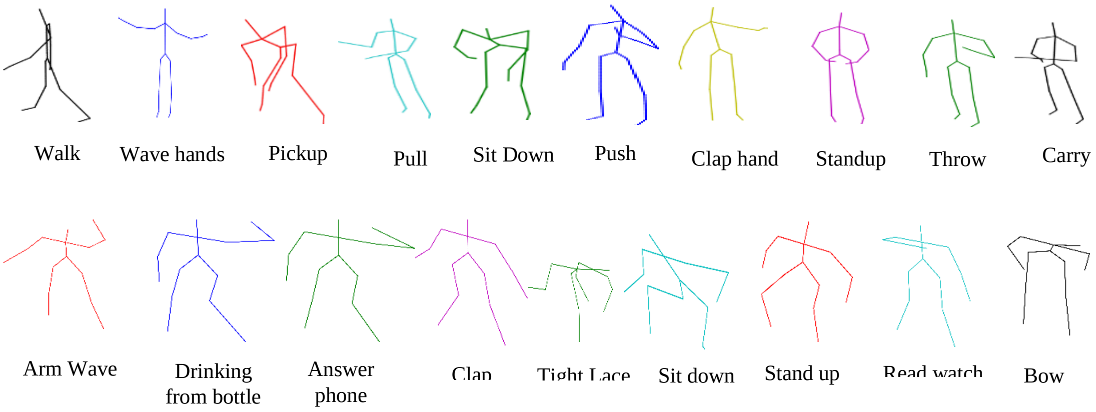


Figure 4.13: Sample frames of (a) UT-Kinect Action 3D Dataset (b) Florence Action 3D dataset

Data augmentation: Deep neural networks demand a large amount of training data to perform efficiently. We have only 557,428, and 200 skeleton sequences, for MSR Action 3D dataset, Florence 3D Action dataset and UT Kinect Action 3D Dataset. Data augmentation helps to reduce the overfitting effect, before processing the data. It includes cropping, horizontal and vertical flip, rotation at 45^0 and -45^0.

Parameter settings: The RIAC-Net architecture is defined and implemented on Python with Keras framework using the Tensor-Flow backend. We used mini-batches of 256 images, and Adam optimizer with default parameters, β1 = 0.9 and β2 = 0.999, during training. The initial learning rate is set to 0.001 and is decreased by a factor of 0.02 after every 20 epochs. The network is trained for each set of part wise input (HS, LH, RH, LL, RL), for 1000 epochs from scratch. To handle the overfitting in the training phase, we adopted weight noise and early stopping [239] along with drop-out strategy.

Table 4.9: List of action classes in three action subsets of MSR Action 3D Dataset

AS1	Horizontal arm wave	Hammer	Forward punch	High Throw	Hand clap	Bend	Tennis serve	Pick up and throw
AS2	Horizontal arm wave	Hand catch	Draw X	Draw Tick	Draw Circle	Two hand wave	Forward kick	Side boxing
AS3	High throw	Forward kick	Side kick	Jogging	Tennis Swing	Tennis Serve	Golf Swing	Pick up and Throw

The experimental results on UT Kinect 3D Action dataset, Florence 3D Action dataset and MSR Action 3D dataset are reported in Table 4.10, 4.11 and 4.12 respectively. The weights $w_i, i\epsilon(1,5)$ corresponding with the best test accuracy achieved are also provided in tables. The Validation loss curves for UT Kinect 3D Action dataset, Florence 3D Action dataset and MSR Action 3D dataset are shown in Fig. 4.14 (a), (b), and (c)-(e), respectively. The Validation losses gradually decrease with the epochs, which confirms the adequate learning of the models for each part. Receiver output Characteristic (ROC) curve plotted between True Positive Rate (TPR) and False Positive Rate (FPR), in Fig. 4.15 (a) and (b), also support the fact that weighted fusion of part-wise skeleton RIAC-Net features turned out better action representation than Full skeletons (FS) for UT Kinect 3D Action dataset, Florence 3D Action dataset and MSR Action 3D dataset. Area Under the Curve (AUC), i.e. $AUC \in (0,1)$, is also computed for each method. The highest AUC values 1.00, 0.97, (0.97,0.99, and 1.00) are obtained for weighted average late fusion strategy over FS and part-wise (HS, LL, LH, RH, RL) skeleton based approaches for UT Kinect 3D Action dataset, Florence 3D Action dataset, MSR Action 3D dataset - AS1, AS2, AS3 action subsets respectively. The achieved accuracy of the proposed work is compared with the other-state-of-arts for UT Kinect 3D Action dataset, Florence 3D Action dataset and MSR Action 3D dataset in Table 4.13, 4.14, and 4.15. The obtained results outperforms many previous studies [240]- [146], [241] - [242]. The proposed work achieved 100% recognition accuracy for UT Kinect 3D Action dataset using Leave-One-Out Cross Validation (LOOCV) scheme. The weighted classification confusion matrix of the UT Kinect 3D action dataset is shown in Fig. 4.16 (a). From where it is clearly evident that each action is recognized correctly without any misclassification irrespective of the presence of high intra class variations and view variations. The obtained result outperforms many previous studies Lie groups [240] , LRCNLG [146] , Grassmann Manifold [243], TS-LSTM [244] which tried to learn geometrical 3D features of human actions using Lie groups, Grassmann Manifold and temporal sliding LSTMs respectively.

The proposed work achieved 98.33% recognition accuracy on Florence 3D

Action dataset, which is 2.96% higher than LRCNLG [146] that integrated Lie groups with deep neural networks to learn the geometrical 3D features. The confusion matrix for Florence 3D action dataset is shown in Fig. 4.16 (b) which shows that very good recognition accuracy is obtained for most of the actions. However, there exist some confusion between similar actions such as ‘Answer Phone’, ‘drink from bottle’, and ‘high arm wave’, ‘stand up’ and ‘sit down’. We have achieved 98.05% fairly a high recognition accuracy on MSR action 3D dataset which outperformed previous works [245] [149]. However, Pham et al. [242] achieved 99.90% accuracy which utilised deep

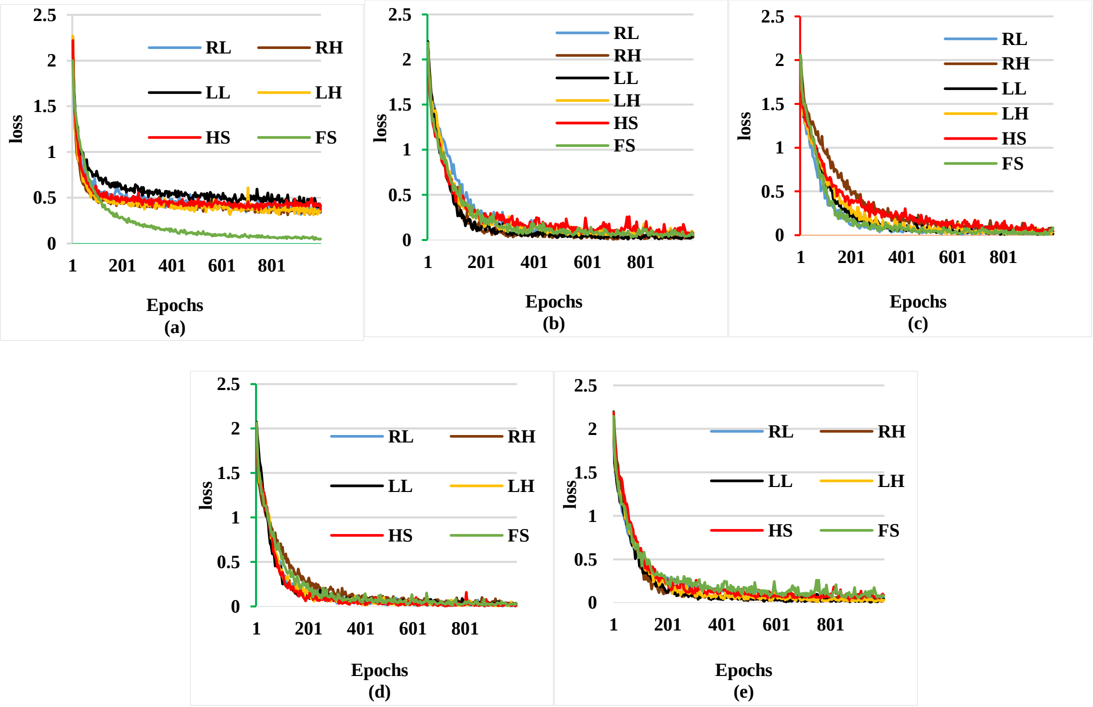
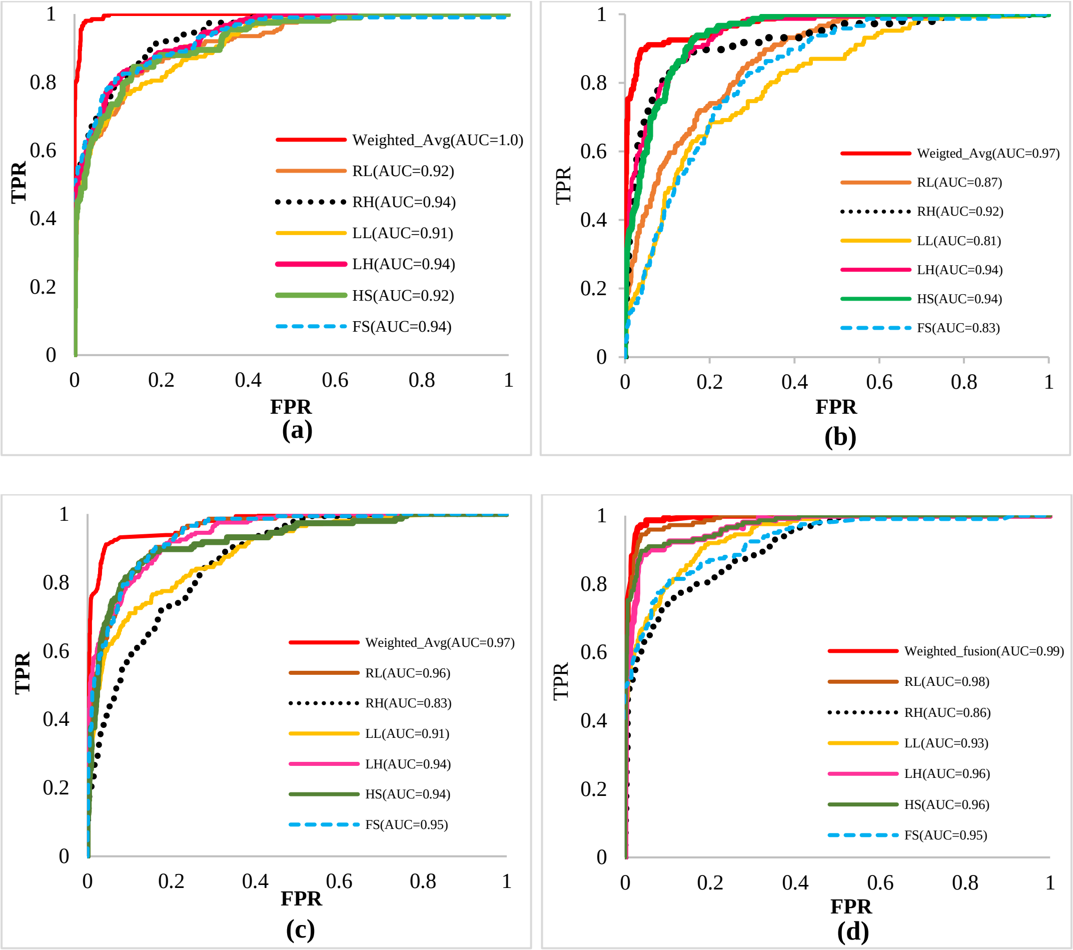


Figure 4.14: Illustration of Part-Wise (RL, RH, LL, LH, HS) and full skeleton (FS) based validation loss curves for (a) UT Kinect 3D Action Dataset, (b) Florence 3D Action dataset, (c)-(e) MSR Action 3D dataset AS1, AS2, and AS3

ResNets to process skeleton data for human action recognition. Some skeleton based methods [149] [245] used pairwise distances between skeleton joints . However, our results obtained on MSR action 3-D dataset show that part wise analysis of whole skeletons followed by late fusion approach is more discriminative approach than taking into consideration the joints separately.

The Classification result for each action subset of the MSR-Action3D dataset are shown in Fig. 4.16 (c), (d), and (e). It is noticed that misclassification occurs only for the actions with high inter class similarity such as ‘draw tick’ and ‘draw *X*’, ‘Pickup and Throw’ and ‘Bend’. Whereas “Forward Kick” and “Tennis Serve” actions which share a large overlap in the sequences, are more challenging to distinguish the two actions in AS3 set.



(a)

(b)

(c)

(d)

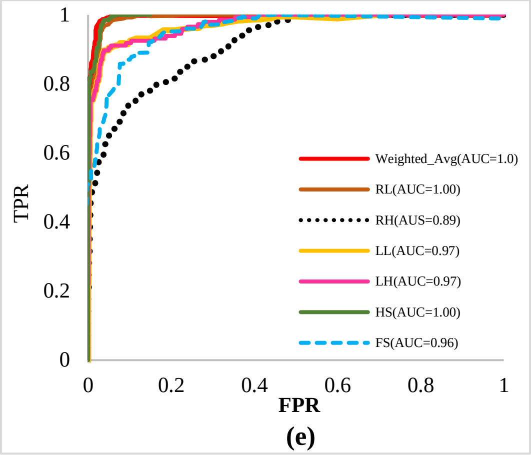


Figure 4.15: ROC Curves of (a) UT Kinect 3D dataset (b) Florence Action 3D dataset (c)-(e) MSR Action 3D dataset: AS1, AS2, AS3

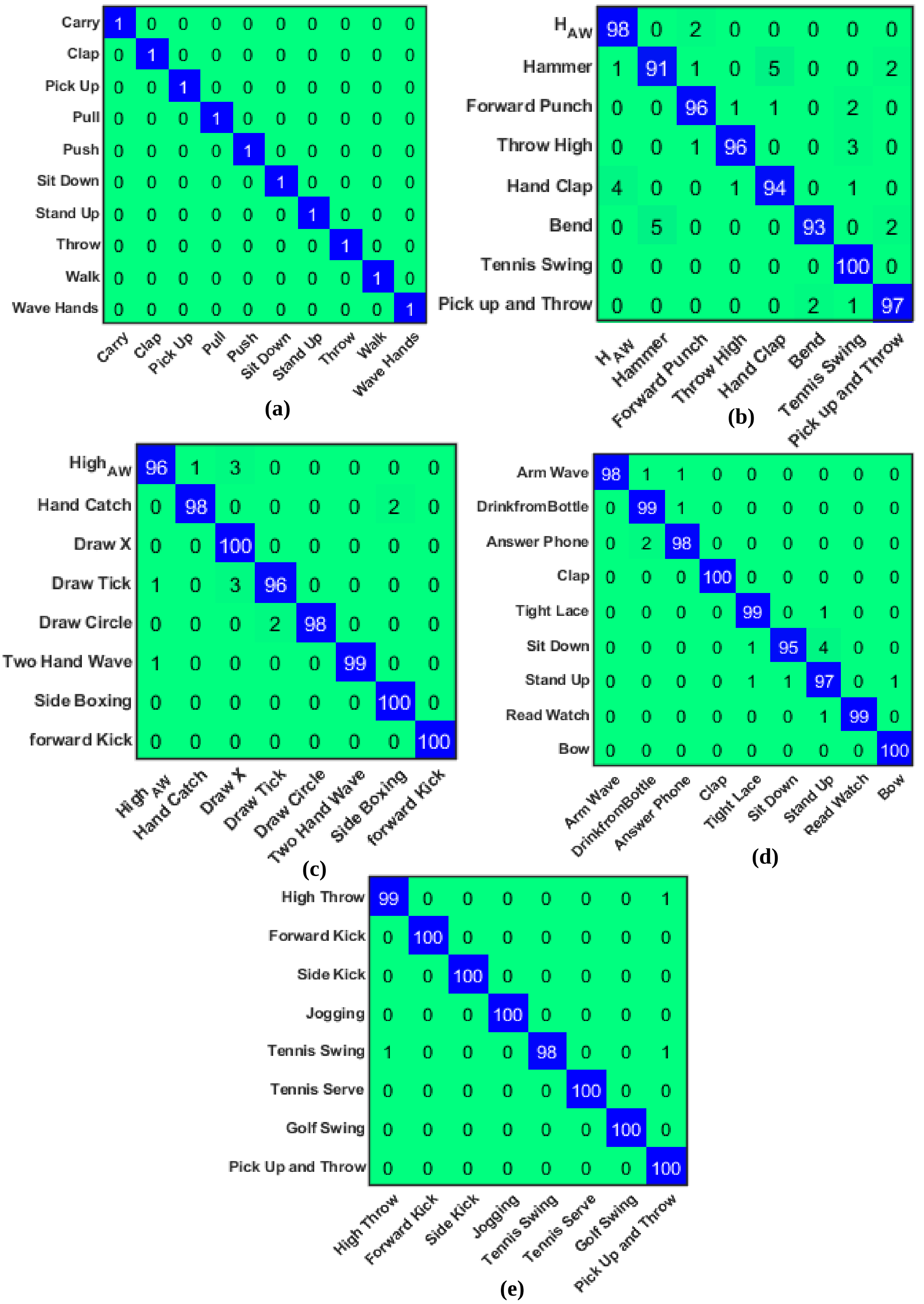
The proposed framework handled this inter class similarity between the two actions and recognized 'Forward Kick' and 'Tennis Serve' with 100% accuracy.

Table 4.10: Performance of the proposed framework for UT Kinect dataset

Parts / Parameters	**FS**	**HS**	**LL**	**RL**	**LH**	**RH**	**Weighted Fusion**
Training Loss	0.2836	0.4480	0.4697	0.3209	0.3164	0.3801	$W_{HS}, W_{LL}, W_{RL}, W_{LH}, W_{RH}$ {2,3,4,4,5}
Training Accuracy	99.96	92.94	92.94	97.65	97.13	95.92	
Test accuracy	97.71	97.49	97.49	96.45	95.94	96.48	**100.00**

Table 4.11: Performance of the proposed framework for Florence 3D Action Dataset

Parts / Parameters	**FS**	**HS**	**LL**	**RL**	**LH**	**RH**	**Weighted Fusion**
Training Loss	0.1221	0.370	0.0279	0.0630	0.0868	0.0127	$\boldsymbol{W_{HS}, W_{LL}, W_{RL}, W_{LH}, W_{RH}}$ **{3,4,2,3,2}**
Training Accuracy	100.00	100.0	100.0	99.69	96.65	100.0	
Test accuracy	95.89	100.00	92.47	95.89	92.63	91.85	**98.33**



(a)

	Carry	Clap	Pick Up	Pull	Push	Sit Down	Stand Up	Throw	Walk	Wave Hands
Carry	1	0	0	0	0	0	0	0	0	0
Clap	0	1	0	0	0	0	0	0	0	0
Pick Up	0	0	1	0	0	0	0	0	0	0
Pull	0	0	0	1	0	0	0	0	0	0
Push	0	0	0	0	1	0	0	0	0	0
Sit Down	0	0	0	0	0	1	0	0	0	0
Stand Up	0	0	0	0	0	0	1	0	0	0
Throw	0	0	0	0	0	0	0	1	0	0
Walk	0	0	0	0	0	0	0	0	1	0
Wave Hands	0	0	0	0	0	0	0	0	0	1

(b)

	H_{AW}	Hammer	Forward Punch	Throw High	Hand Clap	Bend	Tennis Swing	Pick up and Throw
H_{AW}	98	0	2	0	0	0	0	0
Hammer	1	91	1	0	5	0	0	2
Forward Punch	0	0	96	1	1	0	2	0
Throw High	0	0	1	96	0	0	3	0
Hand Clap	4	0	0	1	94	0	1	0
Bend	0	5	0	0	0	93	0	2
Tennis Swing	0	0	0	0	0	0	100	0
Pick up and Throw	0	0	0	0	0	2	1	97

(c)

	High$_{AW}$	Hand Catch	Draw X	Draw Tick	Draw Circle	Two Hand Wave	Side Boxing	forward Kick
High$_{AW}$	96	1	3	0	0	0	0	0
Hand Catch	0	98	0	0	0	0	2	0
Draw X	0	0	100	0	0	0	0	0
Draw Tick	1	0	3	96	0	0	0	0
Draw Circle	0	0	0	2	98	0	0	0
Two Hand Wave	1	0	0	0	0	99	0	0
Side Boxing	0	0	0	0	0	0	100	0
forward Kick	0	0	0	0	0	0	0	100

(d)

	Arm Wave	DrinkfromBottle	Answer Phone	Clap	Tight Lace	Sit Down	Stand Up	Read Watch	Bow
Arm Wave	98	1	1	0	0	0	0	0	0
DrinkfromBottle	0	99	1	0	0	0	0	0	0
Answer Phone	0	2	98	0	0	0	0	0	0
Clap	0	0	0	100	0	0	0	0	0
Tight Lace	0	0	0	0	99	0	1	0	0
Sit Down	0	0	0	0	1	95	4	0	0
Stand Up	0	0	0	0	1	1	97	0	1
Read Watch	0	0	0	0	0	0	1	99	0
Bow	0	0	0	0	0	0	0	0	100

(e)

	High Throw	Forward Kick	Side Kick	Jogging	Tennis Swing	Tennis Serve	Golf Swing	Pick Up and Throw
High Throw	99	0	0	0	0	0	0	1
Forward Kick	0	100	0	0	0	0	0	0
Side Kick	0	0	100	0	0	0	0	0
Jogging	0	0	0	100	0	0	0	0
Tennis Swing	1	0	0	0	98	0	0	1
Tennis Serve	0	0	0	0	0	100	0	0
Golf Swing	0	0	0	0	0	0	100	0
Pick Up and Throw	0	0	0	0	0	0	0	100

Figure 4.16: Confusion Matrix of the a) UT Kinect Dataset, b) Florence Action 3D dataset and MSR action 3D dataset c) AS1, d) AS2 and e) AS3 sets. Where H_A_W: Horizontal Arm Wave and High_A_W: High Arm Wave

Table 4.12: Performance (%) of the proposed framework for MSR Action 3D Dataset under cross-subject evaluation strategy

Dataset subsets	FS	HS	LL	RL	LH	RH	Weighted Fusion Accuracy $W_{HS}, W_{LL}, W_{RL}, W_{LH}, W_{RH}$
AS1	94.8	96	95	98.7	96	90	96.7 { **2,3,3,3,3 }**
AS2	93.3	97.33	100.00	99.5	96.44	93.3	**98.6 { 1,5,4,1,5 }**
AS3	96.5	100	98.2	100	97.3	95.4	**99.9 { 2,4,1,1,3 }**
Overall	94.8	97.77	96.84	99.39	96.58	92.9	**98.40**

Table 4.13: A comparison of the proposed framework with other state-of-the-arts for UT Kinect Action Dataset

Method	Learning Model	Protocol	Accuracy (%)
Feature Combination [246]	K-NN	LOOCV	98
ST-LSTM+Trust Gate [136]	Hierarchal RNN	LOOCV	97
Grassmann Manifold [243]	LTBSVM	LOOCV	88.5
Geometric Features [247]	Multi-layer LSTM	cross validation	95.96
TS-LSTM [244]	Ensemble Temporal sliding LSTM	cross validation	96.97
Lie groups [240]	SVM	One vs all cross validation	97.08
Kernel Linearization [248]	SVM	cross validation	98.2
LRCNLG [146]	LSTM	LOCCV	98.5
Proposed work	**LSTM layers**	**LOOCV**	100.00

Table 4.14: A comparison of the proposed framework with other state-of-the-arts for Florence 3D Action Dataset

Method	Learning Model	Protocol	Accuracy (%)
Lie groups [240]	SVM	One vs all Cross validation	90.88
Kernel Linearization [248]	SVM	Cross validation	95.23
Riemannian Manifold [249]	K-NN	LOOCV	87.04
Mining key pose [250]	Inference Algorithm	LOOCV	92.25
Feature combination [246]	K-NN	LOOCV	94.39
LRCNLG [146]	LSTM	LOOCV	95.37
Proposed work	**LSTM layers**	**LOOCV**	**98.33**

Table 4.15: A comparison of the proposed framework with other state-of-the-arts for MSR Action 3D Dataset cross subject evaluation

Method	Learning Model	AS1	AS2	AS3	Accuracy (%)
HBRNN-L [241]	Hierarchical RNN	93.33	94.64	95.50	94.49
ST-NBNN [149]	Naive-Bayes Nearest-Neighbor	91.50	95.60	97.30	94.80
DMM-LBP-DF [251]	K-ELM	98.10	92.00	94.60	94.90
VBDDM [125]	ProCRC	99.10	92.30	98.20	96.50
Mean3DJ [252]	Random Forest	-	-	-	82.68
Lie group-MinP-PrefixSpan [253]	SVM	-	-	-	97.4
SPMFs [245]	D-CNN	97.06	**99.00**	98.09	**98.05**
ResNet-44 [242]	D-CNN	99.90	99.80	100	99.90
Proposed method	**LSTM layers**	**95.70**	**98.60**	**99.90**	**98.06**

4.3 Significant Outcomes

The outcomes of this chapter are two folds. Firstly, this chapter addresses the problem view variations in action sequences and secondly the high inter class similarity and intra class variations of the action which drastically effect the performance of the automatic human action identification in videos. The experimental results demonstrate some stimulating observations, which are as follows:

- Fusion of multiple features leads to a richer action representation than one feature. Both the proposed deep models for action identification fuses multiple cues that resulted in a highly discriminating action representation which outperforms other existing state-of-the-arts as shown in Table 4.4, 4.5, 4.6, 4.13, 4.14, and 4.15.
- Late fusion generates richer information, however, it demands extensive learning of the deep model for individual features. Transfer learning based feature extraction reduces the computational cost to a great extent.
- It is interestingly observed that RIAC-Net based part wise attention driven CNN

features (local features) weighted fusion result in better action identification than the global features for the complete skeletons.

- Instead of processing each frame of the action, sequence of frames of an action are represented as Compact Action Skeleton Sequence (CASS) which preserves the temporal skeleton joints' details and leads to lesser computations.

This chapter is based on the following works:

- **C. Dhiman**, D.K. Vishwakarma, "View-invariant Deep Architecture for Human Action Recognition using late fusion", *arXiv preprint [online] https://arxiv.org/abs/1912.03632v1*, 2019.
- **C. Dhiman,** D. K. Vishwakarma, P. Aggarwal, "Skeleton based Activity Recognition by Fusing Part-wise Spatio-temporal and Attention Driven Residues". *arXiv preprint [online] https://arxiv.org/abs/1912.00576v1*, 2019.

CHAPTER 5

CONCLUSIONS & FUTURE SCOPE

This chapter highlights the conclusion drawn from this study on the basis of theoretical or experimental contributions made, and the details of future research directions as well as the social and technological impact of the work.

5.1 Conclusions

Four major approaches of the human action identification based on traditional handcrafted features and deep features are presented and these approaches are as follows:

- An abnormal human action identification approach is presented to monitor the daily life actions of elderly people to recognise the abnormal action. The proposed framework encrypts an action in terms of AESI image which is rich in the spatio-temporal details and introduce less computational cost. AESI are processed using $\mathcal{R}$-transform and Zernike moments that introduces translational, scale and rotation invariance with appreciable inter-class separation ability. Henceforth, a complete geometrical transformation invariant feature is obtained with less noise sensitivity. Efficiency of the presented work is validated by using three publically available 3D datasets and our own created dataset. The average recognition accuracy achieved on these datasets are 96.5%, 96.64%, 95.9% and 86.4% on UR fall detection dataset, KARD dataset and new AbHA dataset and multi-view NUCLA dataset, respectively. However, the inter-class similarity between fainting and falling forward needs to be reduced in AbHA dataset.

- An efficient and robust HAR framework is proposed by unifying the Spatial Distribution Gradients (SDGs) and Difference of Gaussian (DoG) based Spatio-temporal interest points (STIP). To handle the illumination variations and recording conditions entropy based texture segmentation is used to extract human silhouette. Spatial Distribution Gradients provide global shape description, which is computed on AEI. AEI represent the entire video sequence temporal and spatial shape variations of the object in a single frame diminishing data storage and computational complexity problems. SDGs is strengthened by scale, rotation, translation and view-invariant property of local STIPs. Which is one of the key reason of obtaining a robust and noise free action modelling and modelling handling view variations efficiently with 95.62% and 89.18% ARA for Ballet and multi-view IXMAS dataset. A codebook of Spatio-temporal interest points is generated per frame for each video sequence and vector quantised code-words are finally used to represent the video sequences. The designed structure for action description is simple yet effective, in terms of time of computation, as observed in Table 3.13, and practical challenge handling capacity.
- A novel two stream RGBD deep framework is proposed that exploits view-invariant characteristics of depth stream and RGB stream. It processes the RGB based motion stream and depth based STD stream independently to exploit the

individual modalities without any influence of each other. Motion stream captures the motion details of the action in the form of RGB-Dynamic images (RGB-DIs) which are processed with fine-tuned InceptionV3 deep network. STD stream captures the view-invariant temporal dynamics of depth frames of key poses using HPM [36] model followed by sequence of Bi-LSTM and LSTM layers that helped to learn long-term view-invariant shape dynamics of the actions. Structural Similarity Index Matrix (SSIM) based key pose extraction helps to inspect only major shape variations during the action reducing the redundant frames having minor shape changes. The late fusion of scores of the motion stream and STD stream helps to make prediction about the action label of the test action sequence. The performance of the framework is validated on three publically available multi-view datasets-NUCLA multi-view dataset, UWA3D II Activity dataset, and NTU RGB-D Activity dataset using cross view and cross-subject cross-validation scheme. The ROC representation of the recognition performance of the proposed framework for each test view exhibits the improved AUC for late fusion over motion and STD streams individually. It is also, noticed that the recognition accuracy of the framework is consistent for different views that confirms the view-invariant characteristics of the framework. In the last, comparisons with other state-of-the-arts are outlined for the proposed deep architecture proving the superiority of the framework in terms of time efficiency and accuracy both.

- An effective skeleton based part-wise spatio-temporal CNN – RIAC Network based 3D human action recognition framework is proposed. It models the dynamics of the action by splitting the skeletons into five parts- Head to Spine (HS), Left Leg (LL), Right Leg (RL), Left Hand (LH), Right Hand (RH). Each part of the skeleton behaves differently for every action which is encrypted using RIAC-Net network which helps to highlight local dynamics {LL, LH, RH, RL, HS} of the action, that proved superior representation than the global action dynamics {FS} of the skeleton. The architecture of the RAIC-Net is designed using the concept of attention based residues and inception blocks. The final action class scores are generated by weighted (decision level) fusion of deep features. The empirical results and the analysis of the performance of our proposed approach exhibit promising results with high accuracies 100%, 98.03%, and 98.7% on UT Kinect Action 3D and Florence 3D actions Dataset and MSR Daily Action3D datasets. Obtained results show that weighted fusion of part wise skeleton action dynamics' learning performs better than FS based action recognition. It is also observed that the proposed model is able to handle the intra class variations and inter class similarity among the actions quite decently.

5.2 Future Research Scope

- The proposed algorithm targeting the abnormal human action identification in video sequences worked satisfactorily with very small computation time. In future, the algorithm can be transformed into real-time normal/abnormal action identification solution to serve healthcare applications systems for elderly care at home.

- The SDG-STIP based hybrid feature vector endows promising results but for multi-textured or in complex textured scenes or cluttered background, a simple entropy based texture segmentation approach may not work that efficiently for silhouette segmentation. Therefore, the silhouette extraction approach needs to be tuned for cluttered background, multi-textured images and occlusion.
- 3D shape analysis of actions can describe the variations in shape geometry of the person more clearly, in comparison to 2D model. Therefore, in the future work, SDG descriptor can be applied on 3D shape models instead of 2D AESI images for better shape description.
- To design a feature vector, codebook is formed using pixel values of the identified DoG based STIPs. However, codebook can be made more effective by introducing STIPs trajectory information.
- A two stream RGBD deep framework performed robustly against view-variations using depth and RGB data. The actual benefit of the framework lies in real time application of the design, which demands lesser storage capacity and computation time for fast identification at application side. Therefore, the predefined inceptionV3 model used for feature extraction from DIs can be replaced by MobileNet or squeezeNet by maintaining an acceptable trade-off between identification performance and number of parameters required.

- The part-wise spatio-temporal attention driven CNN based 3D human action identification framework, utilized only skeletons to develop spatio-temporal features of human poses during the action. Since, depth frames provide richer spatial information of the shapes, integration of both depth and skeleton based human pose representations can be fruitful to develop better and richer action description in order to handle the view variations and occlusion.

5.3 Future Applications

In the future, by utilizing these approaches, one can develop a variety of real life application systems such as:

- Monitoring of public places such as like shopping malls, railways stations, bus stand, parking area is at utmost priority to identify occurrence of any abnormal event due to increase in criminal activities in society. Efficiency of manual monitoring of the area under surveillance, for long hours, decreases with time. Therefore, an automatic surveillance system must be defined which can identify the abnormal actions and generate alarm for the action required against it.

- Nuclear family concept is dominating in urban cities, where the elderly people have to stay alone. In this situation, the health of the elderly people is a serious concern. Because to appoint an assistant in routine with the elder person is not affordable by all. Therefore, keeping in view this fact, a real time efficient human action identification system can play an important role in taking care of health of the elderly people which can detect and notify the occurrence of any abnormal event with the person at home to the close relatives and nearby hospital for necessary action.

- The concept of automatic human actions identification in videos has recently been used to develop a Virtual Exercise Rehabilitation Assistant-VERA which assists the individuals or athletes to measure the efficiency of the exercise done during one session and suggest the correct set of exercise postures/techniques e.g. golf swing, cricket swing etc. Therefore, this field of automatic human action identification is finding new dimension of applications.

- Sports analysis is another important application where automatic human action identification can be used to make unbiased decision about the players which can reduce the incidences of objections raised on umpire's decisions.

- In order to increase the safety and security at home or offices, deployment of an automatic intrusion detection system can be of great help.

CHAPTER 6

REFERENCES

[1] M. Dragone, G. Amato, D. Bacciu, S. Chessa, S. Coleman, M. D. Rocco, C. Gallicchio, C. Gennaro, H. Lozano, L. Maguire, M. McGinnity, A. Micheli, G. M. P. O′Hare, A. Renteria, A. Saffiotti, C. Vairo and P. Vance, "A cognitive robotic ecology approach to self-configuring and evolving AAL systems," *Engineering Applications of Artificial Intelligence,* vol. 45, p. 269–280, 2015.

[2] F. Cardile, G. Iannizzotto and F. L. Rosa, "A vision-based system for elderly patients monitoring," in *3rd International Conference on Human System Interaction*, Rzeszow, 2010.

[3] S. Coşar, G. Donatiello, V. Bogo, C. Garate, L. O. Alvares and F. Brémond, "Toward Abnormal Trajectory and Event Detection in Video Surveillance," *IEEE Transactions on Circuit and Systems for Video Technology,* vol. 27, no. 3, pp. 683-695, 2017.

[4] K. K. Roudposhti , J. Dias, P. Peixoto, V. Metsis and U. Nunes, "A Multilevel Body Motion-Based Human Activity Analysis Methodology," *IEEE Transactions on Cognitive and Developmental Systems,* vol. 9, no. 1, pp. 16-29, 2017.

[5] W. Liu, Y. Fan , T. Lei and Z. Zhang, "Human gesture recognition using orientation segmentation feature on random rorest," in *IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP),* Xi'an, 2014.

[6] M. Stein, H. Janetzko, A. Lamprecht, D. Seebacher, T. Schreck, D. Keim and M. Grossniklaus, "From game events to team tactics: Visual analysis of dangerous situations in multi-match data," in *International Conference on Technology and Innovation in Sports, Health and Wellbeing (TISHW),* Vila Real, 2016.

[7] T. L. Chien, K. L. Su and J. H. Guo, "The multiple interface security robot - WFSR-II," in *IEEE International Safety, Security and Rescue Rototics, Workshop*, Kobe, Japan, 2005.

[8] A. A. Aburomman and M. B. I. Reaz, "Ensemble of binary SVM classifiers based on PCA and LDA feature extraction for intrusion detection," in *Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), IEEE,* Xi'an, China, 2016.

[9] P. Kulkarni, B. Patil and B. Joglekar, " An effective content based video analysis and retrieval using pattern indexing techniques," in *International Conference on Industrial Instrumentation and Control (ICIC),* Pune, 2015.

[10] C.-L. Chou, . H.-T. Chen and S.-Y. Lee, "Multimodal Video-to-Near-Scene Annotation," *IEEE Transactions on Multimedia,* vol. 19, no. 2, pp. 354-366, 2017.

[11] K. Singh , D. K. Vishwakarma, . G. S. Walia and R. Kapoor, “Contrast enhancement via texture region based histogram equalization,” *Journal of Modern Optics,* vol. 63, no. 15, pp. 1444-1450, 2016.

[12] K. G. M. Chathuramali, S. Ramasinghe and R. Rodrigo, “Abnormal Activity Recognition Using Spatio-Temporal Features,” in *7th international conference of Information and Automation of Sustainablilty,* Colombo, 2014.

[13] V. A. Nguyen, T. H. Le and T. T. Nguyen, “Single Camera Based Fall Detection Using Motion and Human shape Features,” in *7th International Symposium on Information and Communication Technology,* Hochiminh city, Vietnam, 2016.

[14] L. Panahi and . V. Ghods, “Human fall detection using machine vision techniques on RGB–D images,” *Biomedical Signal Processing and Control,* vol. 44, pp. 146-153, 2018.

[15] X. Ma, H. Wang, B. Xue, M. Zhou, B. Ji and Y. Li, “Depth-Based Human Fall Detection via Shape Features and Improved Extreme Learning Machine,” *IEEE Journal of biomedical and health informatics,* vol. 18, no. 6, pp. 1915-1922, 2014.

[16] E. Akagunduz, M. Aslan, A. Sengur, H. Wang and M. Ince, “Silhouette Orientation Volumes for Efficient Fall Detection in Depth Videos,” *IEEE Journal of Biomedical and Health Informatics,* vol. PP, no. 99, pp. 2168-2194, 2016.

[17] Z. Zhang , W. Liu, V. Metsis and V. Athitsos, “A Viewpoint-Independent Statistical Method for Fall Detection,” in *21st International Conference on Pattern Recognition* , Tsukuba, 2012.

[18] G. Diraco, A. Leone and P. Siciliano, “An Active Vision System for Fall Detection and Posture Recognition in elderly Healthcare,” in *Design, Automation & Test in Europe Conference & Exhibition,* Dresden, 2010.

[19] N. . Y. Hammerla and T. Plotz, “ Let’s (not) stick together: pairwise similarity biases cross-validation in activity recognition,” in *ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp),* Umeda, Osaka, Japan, 2015.

[20] M. Dragone, G. Amato, D. Bacciu, S. Chessa, S. Coleman, M. D. Rocco, C. Gallicchio, C. Gennaro, H. Lozano, L. Maguire, M. McGinnity, A. Micheli, G. M. P. O'Hare, A. Renteria, A. Saffiotti, C. Vairo and P. Vance, “A cognitive robotic ecology approach to self-configuring and evolving AAL systems,” *Engineering Applications of Artificial Intelligence,* vol. 45, p. 269–280, 2015.

[21] B. Andò, S. Baglio, C. O. Lombardo and V. Marletta, “An Event Polarized Paradigm for ADL Detection in AAL context,” *IEEE Transactions on Instrumentation and Measurement,* vol. 64, no. 7, pp. 1814-1825, 2015.

[22] J. Rafferty, C. D. Nugent, J. Liu and L. Chen, “From Activity Recognition to Intention Recognition for Assisted Living Within Smart Homes,” *IEEE Transactions on Human-Machine Systems,* vol. PP, no. 99, pp. 1-12, 2017.

[23] X. Zhao, A. M. Naguib and S. Lee, “Kinect Based Calling Gesture Recognition for Taking Order Service,” in *23rd IEEE International Symposium on Robot and Human Interactive Communication*, Edinburgh, Scotland, UK, 2014.

[24] J. W. Hsieh, C. H. Chuang, S. Alghyali, H. F. Chiang and C. H. Chiang, “Abnormal Scene Change Detection From a Moving Camera Using Bag of Patches and Spider Web Map,” *IEEE Sensors Journal,* vol. 15, no. 5, pp. 2866-2881, 2015.

[25] J. H. Mosquera, H. Loaiz , S. E. Nope and A. D. Restrepo, “Identifying facial gestures to emulate a mouse: navigation application on Facebook.,” *IEEE Latin America Transactions,* vol. 15, no. 1, pp. 121-128, 2017.

[26] G. Zhu, L. Zhang, P. Shen and J. Song, “Multimodal Gesture Recognition Using 3D Convolution and Convolutional LSTM,” *IEEE Access,* vol. PP, no. 99, pp. 1-1, 2017.

[27] B. Feng, F. He, X. Wang, Y. Wu, H. Wang, S. Yi and W. Liu, “Depth-Projection-Map-Based Bag of Contour Fragments for Robust Hand Gesture Recognition,” *IEEE Transactions on Human-Machine Systems,* vol. PP, no. 99, pp. 1-13, 2016.

[28] G. Zhu , C. Xu , Q. Huang , Y. Rui, S. Jiang , W. Gao and H. Yao, “Event Tactic Analysis Based on Broadcast Sports Video,” *IEEE Transactions on Multimedia,* vol. 11, no. 1, pp. 49-67, 2009.

[29] H.-C. Shih, “A Survey on Content-aware Video Analysis for Sports,” *IEEE Transactions on Circuits and Systems for Video Technology,* vol. PP, no. 99, pp. 1-1, 2017.

[30] S. J. Yu , P. Koh , H. Kwon , D. S. Kim and H. K. Kim, “Hurst Parameter Based Anomaly Detection for Intrusion Detection System,” in *IEEE International Conference on Computer and Information Technology (CIT),* Nadi , 2016.

[31] X. Song and G. Fan, “Joint Key-Frame Extraction and Object Segmentation for Content-Based Video Analysis,” *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 16, no. 7, pp. 904-914, 2006.

[32] J. K. Aggarwal and M. S. Ryoo, “Human Activity Analysis: A Review,” *ACM Computing Surveys*, vol. 43, no. 3, 2011.

[33] H. Permuter, J. Francos and I. Jermyn, “A study of Gaussian mixture models of color and texture features for image classification and segmentation,” *Pattern Recognition,* vol. 39, no. 4, pp. 695-706, 2006.

[34] S. Zeng, R. Huang, Z. Kang and N. Sang, “Image segmentation using spectral clustering of Gaussian mixture models,” *Neurocomputing,* vol. 144, pp. 346-356, 2014.

[35] B. Fernando, E. Gavves, J. Oramas, . A. Ghodrati and T. Tuytelaars, “Modeling video evolution for action recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, Massachusetts, 2015.

[36] H. Rahmani and A. Mian, “3D Action Recognition from Novel Viewpoints,” in *IEEE Conference on Computer Vision and Pattern Recognition,* Las Vegas, NV, USA, 2016.

[37] H. Rahmani, A. Mian and M. Shah, “Learning a Deep Model for Human Action Recognition from Novel Viewpoints,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 40, no. 3, pp. 667-681, 2018.

[38] “EC, “Active ageing special eurobarometer 378,” tech. rep. DG COMM“Research and Speechwriting” Unit, European Comission,” in *Conducted by TNS Opinion & Social at the request of Directorate-General for Employment, Social Affairs and Inclusion*, 2012.

[39] “World Health Organization (WHO) global report on falls prevention in older age,” Geneva, 2008.

[40] S. Zolfaghari and . M. R. Keyvanpour, “SARF: Smart activity recognition framework in Ambient Assisted Living,” in *Federated Conference on Computer Science and Information Systems (FedCSIS),* Gdansk, 2016.

[41] R. Li, B. Lua and K. D. M. Maier, “Cognitive assisted living ambient system: a survey,” *Digital Communications and Networks,* vol. 1, no. 4, pp. 229-252, 2015.

[42] P. Rashidi and A. Mihailidis, “A Survey on Ambient-Assisted Living Tools for Older Adults,” *IEEE Journal of Biomedical and Health Informatics,* vol. 17, no. 3, pp. 579 - 590, 2013.

[43] D. K. Vishwakarma, “Anlysis of Video sequence using Intelligent Techniques,” Delhi Technological University, New Delhi, 2015.

[44] M. Bregonzio, T. Xiang and S. Gong, “Fusing appearance and distribution information of interst points for action recognition,” *pattern Recognition,* vol. 45, no. 3, pp. 1220-1234, 2012.

[45] Y. Zhang, H. Lu, L. Zhang and X. Ruan, “Combining Motion and Appearance Cues for Anomaly Detection,” *Pattern Recognition,* vol. 51, pp. 443-452, 2016.

[46] S. Sedai, M. Bennamoun and D. Q. Huynh, “Discriminative fusion of shape and appearance features for human pose estimation,” *Pattern Recognition,* vol. 46, no. 12, pp. 3223-3237, 2013.

[47] D. Zhao, L. Shao, X. Zhen and Y. Liu, “Combining appearence and structural features for human action recogntion,” *Neurocomputing,* vol. 113, no. 3, pp. 88-96, 2013.

[48] J. L. C. Candás, V. Peláez, G. López, M. Á. Fernández, E. Álvarez and G. Díaz, “An automatic data mining method to detect abnormal humanbehaviour using physical activity measurements,” *Pervasive and Mobile Computing,* vol. 15, pp. 228-241, 2014.

[49] B. Huang, G. Tian , H. Wu and F. Zhou, “A method of abnormal habits recognition in intelligent space,” *Engineering Applications of Artificial Intelligence, ELsevier,* vol. 29, pp. 125-133, 2014.

[50] C. Li, Z. Han, Q. Ye and J. Jiao, “Visual abnormal behavior detection based on trajectory sparse reconstruction analysis,” *Neurocomputing,* vol. 119, pp. 94-100, 2012.

[51] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision,* vol. 60, pp. 91-110, 2008.

[52] J. S. Tham, Y. C. Chang and M. F. A. Fauzi, “Automatic Identification of Drinking Activities at Home using Depth,” in *International Conference on Control, Automation and Information Sciences,* Gwangju,Korea, 2014.

[53] . K. Simonyan and . A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in *Advances in Neural Information Processing Systems (NIPS),* Montreal; Canada, 2014.

[54] O. P. Popoola and K. Wan, “Video-Based Abnormal Human Behavior Recognition—A Review,” *IEEE Transactions on systems, man, and cybernetics,* vol. 42, no. 6, pp. 865-878, 2012.

[55] C. Rougier, J. Meunier, A. St-Arnaud and J. Rousseau, “Robust Video Surveillance for Fall Detection Based on Human Shape Deformation,” *IEEE Transactions on circuits and systems for video technology,* vol. 21, no. 5, pp. 611-622, 2011.

[56] B. Jansen and R. Deklerck, “Context aware inactivity recognition for visual fall detection,” in *Pervasive Health Conference and Workshops, IEEE,* Innsbruck, 2006.

[57] L. Panahi and . V. Ghods, “Human fall detection using machine vision techniques on RGB–D images,” *Biomedical Signal Processing and Control,* vol. 44, pp. 146-153, 2018.

[58] D. 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 Developmental Systems,* pp. 1-1, 2016.

[59] D. K. Vishwakarma and R. Kapoor, “Hybrid classifier based human action recogntion using silhouettes and cells,” *Expert Systems with Applications,* vol. 42, no. 20, pp. 6957-6965, 2015.

[60] S. Al-Ali, M. Milanova and H. . A.-R. Lynn Fox, “Human Action Recognition: Contour-Based and Silhouette-Based Approaches,” *Computer Vision in Control Systems,* vol. 2, pp. 11-47, 2014.

[61] C. Coniglio, C. Meurie, O. Lézoray and M. Berbineau, “People silhouette extraction from people detection bounding boxes in images,” *Pattern Recognition Letters,* vol. 93, pp. 182-191, 2017.

[62] C. Coniglio, C. Meurie, O. Lézoray and M. Berbineau, “A Graph Based People Silhouette Segmentation Using Combined Probabilities Extracted from Appearance, Shape Template Prior, and Color Distributions,” in *International Conference on Advanced Concepts for Intelligent Vision Systems,* Catania, Italy,, 2015.

[63] S. Al-Maadeed, R. Almotaeryi, . R. Jiang and A. Bouridane, “Robust human silhouette extraction with Laplacian fitting,” *Pattern Recognition Letters,* vol. 49, pp. 69-76, 2014.

[64] H. Guo, X. Fan and . S. Wang, “Human attribute recognition by refining attention heat map,” *Pattern Recognition Letters,* vol. 94, pp. 38-45, 2017.

[65] S. Singh, S. Velastin, . H. Ragheb and M. , “A multicamera human action video dataset for the evaluation of action recognition methods,” in *International Conference on Advanced Video and Signal Based Surveillance,* Boston, Massachusetts, 2010.

[66] Z. A. Khan and W. Sohn, “A hierarchical abnormal human activity recognition system based on R-transform and kernel discriminant analysis for elderly health care,” *Computing,* vol. 95, no. 2, pp. 109-127, 2013.

[67] Z. A. Khan and W. Sohn, “Abnormal Human Activity Recognition System based on R-Transform and Kernel Discriminant Technique for Elederly Home Care,” *IEEE Transactions on Consumer Electronics* , vol. 57, no. 4, pp. 1843-1850, 2011.

[68] S. Belongie, J. Malik and . J. Puzicha, “Shape Matching and Object Recognition Using Shape Contexts,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, no. 2, pp. 509-522, 2002.

[69] X. Tao and . Z. Yun, “Fall prediction based on biomechanics equilibrium using Kinect,” *International Journal of Distributed Sensor Networks,* vol. 13, no. 4, 2017.

[70] C. Yuan, X. Li, W. Hu, H. Ling and S. Maybank, “3D R Transform on Spatio-temporal Interest Points for Action Recognition,” in *CVPR,* Portland, OR, USA, 2013.

[71] W. Takano, . Y. Yamada and Y. Nakamur, “Generation of action description from classification of motion and object,” *Robotics and Autonomous Systems,* vol. 91, pp. 247-257, 2017.

[72] F. Patrona, . A. Chatzitofis, D. Zarpalas and P. Daras, “Motion analysis: Action detection, recognition and evaluation based on motion capture data,” *Pattern Recognition,* vol. 76, pp. 612-622, 2018.

[73] X. Wang, C. Qi and F. Lin, "Combined trajectories for action recognition based on saliency detection and motion boundary," *Signal Processing: Image Communication,* vol. 57, pp. 91-102, 2017.

[74] D. . D. Dawn and . S. . H. Shaikh, "A comprehensive survey of human action recognition," *The Visual Computer,* vol. 32, no. 3, pp. 289-306, 2016.

[75] 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.

[76] E. . P. Ijjina and . K. . M. Chalavadi, "Human action recognition in RGB-D videos using motion sequence," *Pattern Recognition,* vol. 72, pp. 504-516, 2017.

[77] H. Aggarwal and D. K. Vishwakarma, "Covariate conscious approach for Gait recognition based upon Zernike moment invariants," *IEEE Transactions on Cognitive and Developmental Systems,* vol. 10, no. 2, pp. 397-407, 2018.

[78] D. K. Vishwakarma, R. Kapoor and A. Dhiman, "Unified framework for human activity recognition: An approach using spatial edge distribution and R-transform," *AEU-International Journal of Electronics and Communications,* vol. 70, no. 3, pp. 341-353, 2016.

[79] S. Gaglio, G. L. Re and M. Morana, "Human Activity Recognition Process Using 3-D Posture Data," *IEEE Transactions on Human-Machine Systems,* vol. 45, no. 5, pp. 586-597, 2015.

[80] I. M. Sintorn and G. Kylberg, "Regional Zernike moments for texture recognition," in *International Conference on Pattern Recognition (ICPR2012),* Tsukuba, 2012.

[81] S. Li, M. C. Lee and C. M. Pun, "Complex Zernike Moments Features for Shape-Based Image Retrieval," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans,* vol. 39, no. 1, pp. 227-237, 2009.

[82] S. Abdelhedi, A. Wali and A. M. Alimi, "Human activity recognition based on mid-level representations in video surveillance applications," in *International Joint Conference on Neural Networks (IJCNN),* Vancouver, BC, 2016.

[83] I. Laptev, "On Space-Time Interest Points," *International Journal of Computer Vision,* vol. 64, no. 2-3, pp. 107-123, 2005.

[84] M. Raptis and S. Soatto, "Tracklet descriptors for action modeling and Video Analysis," in *ECCV, Lecture Notes in Computer Science, vol 6311, Springer*, Berlin, Heidelberg, 2010.

[85] L. Pei, M. Ye, X. Zhao and Y. . D. Bao, "Action recognition by learning temporal slowness invariant features," *The Visual Computer,* vol. 32, no. 11, p. 1395–1404, 2016.

[86] T.-N. Nguyen and K. Miyata, “Multi-scale region perpendicular local binary pattern: an effective feature for interest region description,” *The Visual Computer,* vol. 31, no. 4, pp. 391-406, 2015.

[87] X. Zhu and Z. Liu, “Human behavior clustering for anomaly detection,” *Frontiers of Computer Science in China,* vol. 5, no. 3, pp. 279-289, 2011.

[88] D. Riboni, G. Civitarese and C. Bettini, “Analysis of Long-term Abnormal Behaviors for Early Detection of Cognitive Decline,” in *IEEE International Workshop on PervAsive Technologies and care systems for sustainable Aging-in-place*, Sydney, 2016.

[89] B. Huang, G. Tian , H. Wu and F. Zhou, “A method of abnormal habits recognition in intelligent space,” *Engineering Applications of Artificial Intelligence,* vol. 29, pp. 125-133, 2014.

[90] L. Chen and J. F. H.Wei, “A Survey on human motion analysis using depth imagery,” *Pattern Recognition Letters,* vol. 34, pp. 1995-2006, 2013.

[91] L. L. Presti and M. L. Cascia, “3 D Skeleton based Human Action Classification : A survey,” *Pattern Recognition,* vol. 53, pp. 130-147, 2016.

[92] M. Edwards, J. Deng and X. Xie, “From Pose to Activity : Surveying datasets and introducing CONVERSE,” *Computer Vision and Image Understanding,* vol. 144, pp. 73-105, 2016.

[93] J. Synnott, C. Nugent and P. Jeffers, “Simulation of Smart Home Activity Datasets,” *Sensors,* vol. 15, no. 6, pp. 14162-14179, 2015.

[94] A. Jalal, Y.-H. Kim, . Y.-J. Kim and D. Kim, “Robust human activity recognition from depth video using spatiotemporal multi-fused features,” *Pattern Recognition,* vol. 61, pp. 295-308, 2017.

[95] M. Asadi-Aghbolaghi and S. Kasaei, “Supervised spatio-temporal kernel descriptor for human action recognition from RGB-depth videos,” *Multimedia Tools and Applications,* pp. 1-21, 2017.

[96] L. Yao, W. Min and . K. Lu, “A New Approach to Fall Detection Based on the Human Torso Motion Model,” *Applied Science,* vol. 7, 2017.

[97] C. Rougier, E. Auvinet, J. Rousseau, M. Mignotte and J. Meunier, “Fall Detection from Depth Map Video Sequences,” in *International Conference on Smart Homes and Health Telematics*, Montreal, 2011.

[98] S. Gasparrini, . E. Cippitelli, S. Spinsante and E. Gambi, “A depth-based fall detection system using a Kinect sensor,” *Sensors,* vol. 14, no. 2, pp. 2756-2775, 2014.

[99] L. Yang, Y. Ren, H. Hu and B. Tian, “New fast fall detection method based on spatio-temporal context tracking of head by using depth images,” *Sensors,* vol. 15, pp. 23004-23019, 2015.

[100] B. U. Toreyin, Y. Dedeoglu and A. E. Cetin, "HMM Based Falling Person Detection Using Both Audio and Video," in *Signal Processing and Communications Applications*, Antalya, 2006.

[101] N. Zerrouki, F. Harrou, Y. Sun and A. Houacine, "Accelerometer and Camera-Based Strategy for Improved Human Fall Detection," *Journal of Medical Systems,* vol. 40, no. 12, pp. 1-6, 2016.

[102] R. Nar, A. Singal and P. Kumar, "Abnormal Activity Detection for Bank ATM Surveillance," in *International Conference on Advances in Computing, Communications and Informatics*, Jaipur, India, 2016.

[103] J. Hendryli and M. I. Fanany, "Classifying Abnormal Activities in Exam Using Multi-class Markov Chain LDA Based on MODEC Features," in *Fourth International Conference on Information and Communication Technologies* , Bandung, Indonesia, 2016.

[104] A. A. Chaaraoui, J. R. Padilla-López and F. Flórez-Revuelta, "Abnormal Gait Detection with RGB-D Devices," in *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG),* Ljubljana, Slovenia, 2015.

[105] A. Paiement, L. Tao, M. Camplani, S. Hannuna, D. Damen and M. Mirmehdi, "Online quality assessment of human motion from skeleton data," in *Proceedings of the British Machine Vision Conference*, Nottingham, 2014.

[106] A. Jalal, S. Kamal and D. Kim, "A Depth Video Sensor-Based Life-Logging Human Activity Recognition System for Elderly Care in Smart Indoor environment," *Journal of Sensors,* vol. 14, no. 7, pp. 11735-11759, 2014.

[107] Z. Bian, L. P. Chau and N. M. Thalmann, "Fall detection based on skeleton extraction," in *International Conference on Virtual-Reality Continuum and its Applications in Industry*, Singapore, 2012.

[108] C. Rougier , J. Meunier, A. St-Arnaud and J. Rousseau, "Monocular 3D Head Tracking to Detect Falls of Elderly People," in *IEEE International Conference on Engineering in Medicine and Biology Society,* New York, 2006.

[109] Z. P. Bian, L. P. Chau and N. M. Thalmann, "A Depth Video Approach for Fall Detection Based on Human Joins Height and falling Velocity," in *Proceedings of International Conference on Computer Animation and Social Agents,* Singapore, 2012.

[110] Y. Nizam, M. N. H. Mohd, H. Mohd and M. M. A. Jamil, "Development of Human Fall Detection System using Joint Height, Joint Velocity and Joint Position from Depth Maps," *Journal of Telecommunication, Electronic and Computer Engineering,* vol. 8, no. 6, pp. 125-131, 2016.

[111] Z. P. Bian, L. P. Chau and N. M. Thalmann, "Fall Detection Based on Skeleton Extraction," in *11th International Conference on Virtual-Reality Continuum and its Applications in Industry* , Singapore, 2012.

[112] M. Koohzadi and N. M. Charkari, “Survey on deep learning methods in human action recognition,” *IET Computer Vision,* vol. 11, no. 8, pp. 623-632, 2017.

[113] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, “Gradient-based learning applied to document recognition,” in *Proceedings of IEEE,* 1998.

[114] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg and F. L. Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision,* vol. 115, no. 3, pp. 211-252, 2014.

[115] S. Herath, M. Harandi and F. Porikli, “Going deeper into action recognition: A survey,” *Image and Vision Computing,* vol. 60, pp. 4-21, 2017.

[116] M. Paul , S. Haque and S. Chakraborty, “Human detection in surveillance videos and its applications - A review,” *EURASIP Journal on Advances in Signal Processing,* vol. 1, 2013.

[117] G. W. Taylor, R. Fergus, Y. LeCun and C. Breg, “Convolutional Learning of Spatio-temporal Features,” in *ECCV,* Greece, 2010.

[118] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini and F. Piazza, “Localizing speakers in multiple rooms by using Deep Neural Networks,” *Computer Speech & Language,* vol. 49, pp. 83-106, 2018.

[119] H. Cecotti and A. Graser, “Convolutional Neural Networks for P300 Detection with Application to Brain-Computer Interfaces,” *IEEE Transaction on Pattern Analysis and Machine Intelligence,* vol. 33, no. 3, pp. 433-445, 2010.

[120] W. Yin, K. Kann, M. Yu and H. Schütz, “Comparative Study of CNN and RNN for Natural Language Processing,” *arXiv:1702.01923,* 2017.

[121] G. Iannizzotto, P. Lanzafame and F. L. Rosa, “A CNN-based framework for 2D still-image segmentation,” in *International Workshop on Computer Architecture for Machine Perception,* Palermo, Italy, 2005.

[122] L. Wang, Y. Xu, J. Cheng , H. Xia and J. Yin, “Human Action Recognition by Learning Spatio-Temporal Features With Deep Neural Network,” *IEEE Access,* vol. 6, pp. 17913-17922, 2018.

[123] M. Z. Uddin, W. haksar and J. Torresen, “Facial Expression Recognition Using Salient Features and Convolutional Neural Network,” *IEEE Access,* vol. 5, p. IEEE Access, 2017.

[124] 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.

[125] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems (NIPS)*, Montreal; Canada, 2014.

[126] L. Jing, Y. Ye, X. Yang and Y. Tian, "3D convolutional neural network with multi-model framework for action recognition," in *International Conference on Image Processing (ICIP)*, Beijing, China, 2017.

[127] Z. Liua, C. Zhangb and Y. Tian, "3D-based Deep Convolutional Neural Network for action recognition with depth sequences," *Image and Vision Computing,* vol. 55, no. 2, pp. 93-100, 2016.

[128] J. Liu, G. Wang, L. Y. Duan, K. Abdiyeva and A. C. Kot, "Skeleton-Based Human Action Recognition With Global Context-Aware Attention LSTM Networks," *IEEE Transactions on Image Processing,* vol. 27, no. 4, pp. 1586-1599, 2018.

[129] Y. Yang , I. Saleemi and M. Shah , "Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions," *IEEE Transaction on Pattern Analysis and Machine Intelligence,* vol. 35, pp. 1635-1648, 2013.

[130] A. Diba, M. Fayyaz, V. Sharma, A. H. Karami, M. M. Arzani, R. Yousefzadeh and L. V. Gool, "Temporal 3D ConvNets: New Architecture and Transfer Learning for Video Classification," vol. arXiv: 1711.08200, 2017.

[131] D. K. Vishwakarma and K. Singh, "Human Activity Recognition Based on Spatial Distribution of Gradients at Sublevels of Average Energy Silhouette Images," *IEEE Transactions on Cognitive and Developmental Systems,* vol. 9, no. 4, pp. 316-327, 2017.

[132] Z. Tu, W. Xie, Q. Qin, R. Poppe, R. C. Veltkamp, B. Li and J. Yuanf, "Multi-stream CNN: Learning representations based on human-related regions for action recognition," *Pattern Recognition,* vol. 79, pp. 32-43, 2018.

[133] H. Wang, Y. Yang, E. Yang and C. Deng, "Exploring hybrid spatio-temporal convolutional networks for human action recognition," *Multimedia Tools and Applications,* vol. 76, pp. 15065-15081, 2017.

[134] Y. Liu, Q. Wu, L. Tan and H. shi, "Gaze-Assisted Multi-Stream Deep Neural Network for Action Recognition," *IEEE Access,* vol. 5, pp. 19432-19441, 2017.

[135] W. Lin, Y. Mi, J. Wu, K. Lu and H. Xiong, "Action Recognition with Coarse-to-Fine Deep Feature Integration and Asynchronous Fusion," *arXiv:1711.07430* , 2018.

[136] J. Liu, N. Akhtar and A. M. Saeed , "Viewpoint Invariant RGB-D Human Action Recognition," in *International Conference on Digital Image Computing: Techniques and Applications*, Sydney, 2017.

[137] Y. Du, W. Wang and L. Wang, “Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, 2015.

[138] A. Shahroudy, J. Liu, T.-T. Ng and G. Wang, “NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016.

[139] W. Zhu , C. Lan , J. Xing, W. Zen, Y. Li, L. Shen and X. Xie, “Cooccurrence Feature Learning for Skeleton based Action Recognition using Regularized Deep LSTM Networks,” in *30th AAAI Conference on Artificial Intelligence (AAAI)*, Phoenix, Arizona USA, 2016.

[140] Y. Hou , Z. Li, P. Wang and W. Li , “skeleton Optical Spectra Based Action Recognition Using Convolutional Neural Networks,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 3, pp. 807-811, 2018.

[141] S. Song', C. Lan , J. Xing , W. Zeng and J. Liu , “An End-to-End SpatioTemporal Attention Model for Human Action Recognition from Skeleton data,” in *Thirty-First AAAI Conference on Artificial Intelligence*, California, USA, 2017.

[142] T. S. Kim and A. Reiter, “Interpretable 3D Human Action Analysis with Temporal Convolutional Networks,” *arXiv preprint arXiv:1704.04516*, 2017.

[143] Z. Shi and T.-K. Kim, “Learning and Refining of Privileged Information-based RNNs for Action,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, Hawaii, 2017.

[144] B. Mahasseni and S. Todorovic, “Regularizing Long Short Term Memory with 3D Human-Skeleton Sequences for Action Recognition,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, United States, 2017.

[145] J. Tu, M. Liu and H. Liu, “Skeleton based Human Action Recognition Using Spatial Temporal 3D Convolutional Neural Network,” in *IEEE International Conference on Multimedia and Expo*, San Diego, CA, USA, 2018.

[146] M. Liu , H. Liu and C. Chen, “Enhanced skeleton visualization for view invariant human action recognition,” *Pattern Recognition*, vol. 68, pp. 346-362, 2017.

[147] M. Rhif, H. Wannous and I. R. Farah, “Action Recognition from 3D Skeleton Sequences using Deep Networks on Lie Group Features,” in *International Conference on Pattern Recognition (ICPR)*, Beijing, China, 2018.

[148] H. Chen, G. Wang, J.-H. Xue and L. He, “A novel hierarchical framework for human action recognition,” *Pattern Recognition*, vol. 55, pp. 148-159, 2016.

[149] B. B. Amor, J. Su and A. Srivastava, “Action recognition using rate-invariant analysis of skeletal shape trajectories,” *IEEE Transaction on Pattern Analysis and Machine Intelligence (TPAMI),* vol. 38, no. 1, pp. 1-13, 2016.

[150] J. Weng , C. Weng and J. Yuan, “SpatioTemporal Naive-Bayes Nearest-Neighbor (ST-NBNN) for skeleton-based action recognition,” in *Conference on Computer Vision and Pattern Recognition (CVPR),* Hawaii, 2017.

[151] F. Ofli , R. Chaudhry , G. Kurillo , R. Vidal and R. Bajcsy, “Sequence of the Most Informative Joints (SMIJ): A new representation for human skeletal action recognition,” *Journal of Visual Communication and Image Representation,* vol. 25, no. 1, pp. 24-38, 2014.

[152] C. Wang, Y. Wang and A. L. Yuille, “An approach to pose-based action recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Portland, Oregon, 2013.

[153] I. Lillo, J. C. Niebles and A. Soto, “A Hierarchical Pose-Based Approach to Complex Action Understanding Using Dictionaries of Actionlets and Motion Poselets,” in *Conference on Computer Vision and Pattern Recognition (CVPR),* Nevada, United States, 2016.

[154] S. Tabbone, L. Wendling and J. -P. Salmon, “A new shape descriptor defined on the Radon transform,” *Computer Vision and Image Understanding,* vol. 102, no. 1, pp. 42-51, 2006.

[155] F.Zernike, “Beugungstheorie des schneidenver-fahrens und seiner,” in *Physica*, vol. 1, 1934, pp. 689-704.

[156] S. Jin, . S. Li, P. Sun and W. Cai, “Zernike moment based time sensitive targets detection and recognition,” in *IEEE Chinese Guidance, Navigation and Control Conference,* Nanjing, 2016.

[157] B. Kwolek and M. Kepski, “Human fall detection on embedded platform using depth maps and wireless accelerometer,” *Computer Methods and Programs in Biomedicine,* vol. 117, no. 3, pp. 489-501, 2014.

[158] N. Zerrouki and A. Houacine, “Combined curvelets and hidden Markov models for human fall detection,” *Multimedia Tools and Applications,* vol. 77, pp. 6405-6424, 2018.

[159] Y. Yun and I. Y.-H. Gu, “Human fall detection via shape analysis on Riemannian manifolds with applications to elderly care,” in *IEEE International Conference on Image Processing,* Quebec City, QC, Canada, 2015.

[160] G. Goudelis, G. Tsatiris, K. Karpouzis and S. Kollias, “Fall detection using History Triple Features,” in *Proceedings of the 8th ACM International Conference on Pervasive Technologies Related to Assistive Environments,* Corfu, Greece, 2015.

[161] A. N. Marcos, G. Azkune and I. C. Arganda, “Vision-Based Fall Detection with Convolutional Neural Networks,” *Wireless Communications and Mobile Computing,* 2017.

[162] E. Cippitelli, S. Gasparrini, E. Gambi and S. Spinsante, “A Human Activity Recognition System Using Skeleton Data from RGBD Sensors,” *Computational Intelligence and Neuroscience,* vol. 3, pp. 1-14, 2016.

[163] D. K. Vishwakarma, “Datasets,” Delhi Technological University, 2018. [Online]. Available: http://www.dtu.ac.in/Web/Departments/InformationTechnology/faculty/dkvishwakarma.php.

[164] N. E. D. E. Madany, Y. He and L. Guan, “Integrating Entropy Skeleton Motion Maps and Convolutional Neural Networks for Human Action Recognition,” in *International Conference on Multimedia and Expo* , San Diego, CA, USA, 2018.

[165] H. H. Pham, L. Khoudoura, P. Zegersc and S. A. Velastin, “Exploiting deep residual networks for human action recognition from skeletal data,” *Computer Vision and Image Understanding,* 2018.

[166] J. Wang, B. X. Nie, Y. Xia, Y. Wu and S. C. Zhu, “Cross-view action modeling, learning and recognition,” in *CVPR*, Columbus, Ohio, 2014.

[167] J. Liu and A. S. Mian, “Learning Human Pose Models from Synthesized Data for Robust RGB-D Action,” *CoRR,* vol. arXiv:1707.00823v2, 2017.

[168] R. Li and . T. Zickler, “Discriminative virtual views for crossview action recognition,” in *International Conference on Computer Vision and Pattern Recognition*, Providence, 2012.

[169] Z. Zhang, . C. Wang, B. Xiao, . W. Zhou, S. Liu and C. Shi, “Cross-view action recognition via a continuous virtual path,” in *CVPR,* 2013.

[170] H. Rahmani and A. Mian, “Learning a non-linear knowledge transfer model for cross-view action recognition,” in *IEEE Conference on Computer Visoin and Pattern Recogntion*, Boston, Massachusetts, 2015.

[171] J. Liu, N. Akhtar and . A. Mian, “Skepxels: Spatio-temporal Image Representation of Human Skeleton Joints for Action Recognition,” *CoRR,* vol. arXiv:1711.05941v4, 2018.

[172] D. K. Vishwakarma, R. Kapoor, R. Maheshwari, V. Kapoor and S. Raman, “Recognition of Abnormal Human Activity Using the Changes in Orientation of Silhouette in Key Frames,” in *2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, 2015.

[173] D. K. Vishwakarma, R. Kapoor and A. Dhiman, “Unified framework for human activity recognition: An approach using spatial edge distribution and R-transform,” *International Journal of Electronics and Communications,* vol. 70, no. 3, pp. 341-353, 2016.

[174] S. Brutzer , B. Höferlin and G. Heidemann, "Evaluation of background subtraction techniques for video surveillance," in *IEEE conference on computer vision and pattern recognition (CVPR)*, Colorado Springs, CO, USA, 2011.

[175] T. Ojala and M. Pietikainen, "Unsupervised texture segmentation using feature distributions," *Pattern Recognition,* vol. 32, no. 3, pp. 477-486, 1999.

[176] M. Heikkila and M. Pietikainen, "A Texture-Based Method for Modeling the Background and Detecting Moving Objects," *IEEE Transaction in Pattern Analysis and Machine Intelligence,* vol. 28, no. 4, pp. 657-662, 2006.

[177] A. Rampun, H. Strange and R. Zwiggelaar, "Texture Segmentation Using Different Orientations of GLCM Features," in *International Conference on Computer Vision*, Germany, 2013.

[178] R. M. Haralick, K. Shanmugam and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems Man, and Cybernetics,* vol. SMC, no. 6, pp. 610-621, 1973.

[179] L. Soh and C. Tsatsoulis, "Texture analysis of sar sea ice imagery using gray level co-occurrence matrices," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 37, no. 2, pp. 780-795, 1999.

[180] M. Komorkiewicz and M. Gorgon, "Foreground object features extraction with GLCM texture descriptor in FPGA," in *IEEE conference on design and architectures for signal and image processing (DASIP)*, Cagliari, Italy, 2013.

[181] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), New York: Springer-Verlag, 2006.

[182] S. C. Johnson, "Hierarichal Clustering Schemes," *Pyschometrica,* vol. 32, no. 3, pp. 241-254, 1967.

[183] A. Y. Ng, A. I. Jordan and Y. Weiss, "On spectral clustering : Analysis and an algorithm," in *NIPS*, 2001.

[184] C. Schuldt, I. Laptev and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proceedings of the 17th International Conference on Pattern Recognition*, Cambridge, UK , 2004.

[185] L. Gorelick, M. Blank, E. Shechtman, M. Irani and R. Basri, "Actions as space–time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247-2253, 2007.

[186] 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.

[187] D. Weinland, . R. Ronfard and E. Boyer, “Free viewpoint action recognition using motion history volumes,” *Computer Vision and Image Understaning,* vol. 104, no. 2-3, pp. 249-257, 2006.

[188] H. Rahmani and . A. Mian, “3D Action Recognition from Novel Viewpoints,” in *CVPR,* Las Vegas, 2016.

[189] *CMU motion capture database, http://mocap.cs..*

[190] L. Liu, L. Shao, X. Li and K. Lu, “Learning Spatio-Temporal Representations for Action Recognition: A Genetic Programming Approach,” *IEEE Transactions on cybernetics,* vol. 46, no. 1, pp. 158-170, 2016.

[191] A. A. Chaaraoui, P. C. Pérez and F. F. Revuelta, “Sihouette-based human action recognition using sequences of key poses,” *Pattern Recognition Letters,* vol. 34, no. 15, pp. 1799-1807, 2013.

[192] D. Wu and L. Shao, “Silhouette Analysis-Based Action Recognition Via Exploiting Human Poses,” *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 23, no. 2, pp. 236-243, 2013.

[193] G. Goudelis, K. Karpouzis and S. Kollias, “Exploring trace transform for robust human action recognition,” *Pattern Recognition,* vol. 46, no. 12, pp. 3238-3248, 2013.

[194] Y. Fu, T. Zhang and W. Wang, “Sparse coding-based space-time video representation for action recognition,” *Multimedia Tools and Applications,* vol. 76, no. 10, pp. 12645-12658, 2017.

[195] H. Han and X. J. Li, “Human action recognition with sparse geometric features,” *The Imaging Science Journal,* vol. 63, pp. 45-53, 2015.

[196] J. Lei, G. Li, J. Zhang, Q. Guo and D. Tu, “Continuous action segmentation and recognition using hybrid convolutional neural network-hidden Markov model model,” *IET Computer Vision,* vol. 10, no. 6, pp. 537-544, 2016.

[197] H. Liu, N. Shu, Q. Tang and W. Zhang, “Computational Model Based on Neural Network of Visual Cortex for Human Action Recognition,” *IEEE Transaction on neural networks and learning systems,* vol. PP, no. 99, pp. 1-14, 2017.

[198] S. Sadek, A. A. Hamadi, M. Elmezain, B. Michaelis and U. Sayed, “Human Action Recognition via Affine Moment Invariants,” in *International conference on Pattern Recognition* , Tsukuba, Japan , 2012.

[199] B. Saghafi and D. Rajan, “Human action recognition using Pose-based disriminant embedding,” *Signal Processing: Image Communication,* vol. 27, no. 1, pp. 96-111, 2012.

[200] S. A. Rahman, I. Song, M. K. H. Leung, I. Lee and K. Lee, "Fast action recognition using negative space features," *Expert Systems with Applications,* vol. 41, no. 2, pp. 574-587, 2014.

[201] B. Li, O. I. Camps and M. Sznaier, "Cross-view Activity Recognition using Hankelets," in *IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, 2012.

[202] Y. Shi , Y. Tian , Y. Wang and T. Huang, "Sequential Deep Trajectory Descriptor for Action Recognition With Three-Stream CNN," *IEEE Transactions on Multimedia,* vol. 19, no. 7, pp. 1510-1520, 2017.

[203] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *IEEE conference on Computer Vision and Pattern Recognition, CVPR*, Anchorage, AK, USA , 2008.

[204] 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.

[205] X. L. Ming, H. J. Xia and T. L. Zheng, "Human action recognition based on chaotic invariants," *J. South Cent. Univ.,* vol. 20, pp. 3171-3179, 2014.

[206] 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.

[207] X. Wu, D. Xu, L. Duan and J. Luo, "Action recognition using context and appearance distribution features," in *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, Providence, RI, 2011.

[208] D. Weinland, M. Özuysal and P. Fua , "Making Action Recognition Robust to Occlusions and Viewpoint Changes," in *In Proceedings of the European Conference on Computer Vision (ECCV),* Crete, Greece, 2010.

[209] X. Wu and Y. Jia, "View-Invariant Action Recognition Using Latent Kernelized Structural SVM," in *Proceedings of the 12th European conference on Computer Vision (ECCV),* Florence, Italy, 2012.

[210] J. Wang, H. Zheng, J. Gao and J. Cen, "Cross-View Action Recognition Based on a Statistical Translation Framework," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 26, no. 8, pp. 1461-1475, 2016.

[211] E. A. Mosabbeb, K. Raahemifar and M. Fathy, "Multi-View Human Activity Recognition in Distributed Camera," *Sensors,* vol. 13, no. 7, pp. 8750-8770, 2013.

[212] . A. Krizhevsky, . I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Neural Information Processing Systems (NIPS),* Nevada, 2012.

[213] W. Zhou, A. C. Bovik , H. R. Sheikh and E. P. Simoncelli, "Image Qualifty Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing,* vol. 13, no. 4, pp. 600-612, 2004.

[214] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transaction on Pattern Analysis and Machine Intelligence,* vol. 23, no. 3, pp. 257-267, 2011.

[215] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing,* vol. 14, no. 3, pp. 199-222, 2004.

[216] H. Bilen, . B. Fernando, E. Gavves, A. Vedaldi and S. Gould, "Dynamic Image Networks for Action Recognition," in *IEEE Conference on Computer Visoin and Pattern Recogntion,* Las Vegas, 2016.

[217] J. Wang, B. X. Nie, Y. Xia, Y. Wu and S. C. Zhu, "Cross-view action modeling, learning and recognition," in *IEEE Conference on Computer Visoin and Pattern Recogntion,* Columbus, Ohio, 2014.

[218] H. Rahmani, A. Mahmood, D. Huynh and A. Mian, "Histogram of oriented principal components for cross-view action recognition," *IEEE Transaction on Pattern Analysis and Machine Intelligence,* vol. 38, no. 12, pp. 2430-2443, 2016.

[219] A. Shahroudy, J. Liu, T. T. Ng and G. Wang, "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis," in *Conference on Computer Vision and Pattern Recognition,* Nevada, United States, 2016.

[220] 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.

[221] Z. Zhang, . C. Wang, B. Xiao, . W. Zhou, S. Liu and C. Shi, "Cross-view action recognition via a continuous virtual path," in *IEEE Conference on Computer Visoin and Pattern Recogntion,* 2013.

[222] A. Gupta, J. Martinez, J. J. Little and R. J. Woodham, "3D Pose from Motion for Cross-View Action Recognition via Non-linear Circulant Temporal Encoding," in *IEEE Conference on Computer Visoin and Pattern Recogntion,* Columbus, 2014.

[223] H. Wang, A. Kläser, C. Schmid and C. L. Liu, "Action recognition by dense trajectories," in *IEEE Conference on Computer Visoin and Pattern Recogntion*, Providence, RI, 2011.

[224] S. Song, C. Lan, J. Xing, W. Zeng and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *AAAI conference on Artificial Intelligence*, Califirnia, USA, 2017.

[225] J. Liu, A. Shahroudy, . D. Xu and G. Wang, “Spatio-temporal LSTM with trust gates for 3D human action recognition,” in *European Conference on Computer Vision (ECCV),* Amsterdam, 2016.

[226] T. S. Kim and A. Reiter, “Interpretable 3D Human Action Analysis with Temporal Convolutional,” in *Conference on Computer Vision and Pattern Recognition workshop,* Honolulu, Hawaii, 2017.

[227] A. S. Keçeli, “Viewpoint projection based deep feature learning for single and dyadic action recognition,” *Expert Systems With Applications,* vol. 104, pp. 235-243, 2018.

[228] H.-H. Phama, L. Khoudoura, A. Crouzil, P. Zegers and S. A. Velastin, “Exploiting deep residual networks for human action recognition from skeletal data,” *Computer Vision and Image Understanding,* vol. 170, pp. 51-66, 2018.

[229] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, “Going deeper with convolutions,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015.

[230] K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition,” *arXiv preprint arXiv:1512.03385,* 2015.

[231] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, “Attention Is All You Need,” in *31st Conference on Neural Information Processing Systems (NIPS),* Long Beach, CA, USA, 2017.

[232] K. He , X. Zhang, S. Ren and J. Sun, “Identity mappings in deep residual networks,” *arXiv preprint arXiv:1603.05027,* 2016.

[233] E. Shelhamer, J. Long and T. Darrell, “Fully convolutional networks for semantic segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence* , vol. 39, no. 4, pp. 640-651, 2017.

[234] D. Bahdanau, K. Cho and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv:1409.0473,* 2014.

[235] D. Britz, A. Goldie, M. T. Luong and Q. Le, “Massive exploration of neural machine translation architectures,” *arXiv:1703.03906,* 2017.

[236] C. G. Snoek, M. Worring and A. W. Smeulders, “Early versus Late Fusion in Semantic Video Analysis,” in *Thirteenth ACM International Conference on Multimedia,* Singapore, 2005.

[237] L. Xia , C. C. Chen and J. K. Aggarwal, “View invariant human action recognition using histograms of 3D joints,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops,* Providence, RI, USA, 2012.

[238] 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 Conference on Computer Vision and Pattern Recognition Workshops*, Portland, OR, USA, 2013.

[239] W. Li , Z. Zhang and Z. Liu, “Action recognition based on a bag of 3d points,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, San Francisco, CA, USA, 2010.

[240] A. Graves, “Practical variational inference for neural networks,” in *International Conference on Neural Information Processing Systems*, Granada, Spain , 2011.

[241] R. Vemulapalli , F. Arrate and R. Chellappa, “Human action recognition by representing 3d skeletons as points in a lie group,” in *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014.

[242] Y. Du, Y. Fu and L. Wang, “Representation Learning of Temporal Dynamics for Skeleton-Based Action Recognition,” *IEEE Transactions on Image Processing,* vol. 25, no. 7, pp. 3010-3022, 2016.

[243] H. H. Pham, L. Khoudour, A. Crouzil and P. Zegers, “Exploiting deep residual networks for human action recognition from skeletal data,” *Computer Vision and Image Understanding,* 2018.

[244] R. Slama, H. Wannous, M. Daoudi and A. Srivastava, “Accurate 3D action recognition using learning on the Grassmann manifold,” *Pattern Recognition,* vol. 48, no. 2, pp. 556-567, 2015.

[245] I. Lee, D. Kim , S. Kang and S. Lee, “Ensemble Deep Learning for Skeleton-Based Action Recognition Using Temporal Sliding LSTM Networks,” in *IEEE International Conference on Computer Vision (ICCV),* Venice, Italy, 2017.

[246] H. H. Pham, L. Khoudoury, A. Crouzil, P. Zegers and S. A. Velastiny, “Skeletal movement to color map: A novel representation for 3D Action Recognition with Inception Residual Networks,” *arXiv:1807.07033v1 [cs.CV]* , 2018.

[247] D. C. Luvizon, . H. Tabia and D. Picard, “Learning features combination for human action recognition from skeleton sequences,” *Pattern Recognition Letters,* vol. 99, no. 1, pp. 13-20, 2017.

[248] 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, USA, 2017.

[249] P. Koniusz, A. Cherian and F. Porikli, “Tensor representations via kernel linearization for action recognition from 3d skeletons,” in *European Conference on Computer Vision (ECCV),* Amsterdam, The Netherlands, 2016.

[250] M. Devanne, H. Wannous , S. Berretti, P. Pala , M. Daoudi 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.

[251] C. Wang , Y. Wang and A. L. Yuille, “Mining 3D Key-Pose-Motifs for Action Recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016.

[252] C. Chen , R. Jafari and N. Kehtarnavaz, “Action recognition from depth sequences using depth motion maps-based local binary patterns,” in *Winter Conference on Applications of Computer Vision (WACV),* Hawaii, 2015.

[253] A. Tamou , L. Ballihi and D. Aboutajdine , “Automatic Learning of Articulated Skeletons based on Mean of 3D Joints for Efficient Action Recognition,” *International Journal of Pattern Recognition and Artificial Intelligence,* vol. 31, no. 4, pp. 1-17, 2017.

[254] G. Li , K. Liu , W. Ding , F. Cheng and B. Chen, “Key Skeleton Pattern Mining on 3D Skeletons Represented by Lie Group for Action Recognition,” *Mathematical Problems in Engineering,* vol. 2018, 2018.

AUTHOR BIOGRAPHY

	Chhavi Dhiman 2K16/Ph.D./EC/07 Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India Email: chhavi1990delhi@gmail.com

Chhavi Dhiman received the Bachelor of Technology (B.Tech.) from Indira Gandhi Delhi Technological University for Women (IGDTUW) (formerly known as Indira Gandhi Institute of Technology, GGSIPU), Delhi, India, in the year 2011 and the Master of Technology (M.Tech.) from Delhi Technological University (DTU), Delhi, India in the year 2014. Presently, she is working as a PhD scholar in the Department of Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India-110042. Her research interests include computer vision, Pattern Recognition, Image Processing, Deep Learning and Machine Learning.