ANALYSIS OF VIDEO SEQUENCES USING INTELLIGENT TECHNIQUES

A THESIS

SUBMITTED TO THE DELHI TECHNOLOGICAL UNIVERSITY

FOR THE AWARD OF THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

Electronics and Communication Engineering

SUBMITTED BY

DINESH KUMAR VISHWAKARMA

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

DELHI- 110042 (INDIA)

OCTOBER-2015

ANALYSIS OF VIDEO SEQUENCES USING INTELLIGENT TECHNIQUES

BY

DINESH KUMAR VISHWAKARMA

SUBMITTED TO THE DELHI TECHNOLOGICAL UNIVERSITY IN PARTIAL

FULFILMENTS OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

ELECTRONICS AND COMMUNICATION ENGINEERING

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
DELHI- 110042 (INDIA)
OCTOBER-2015

CERTIFICATE

This is to certify that the thesis entitled **"Analysis of video sequences using intelligent techniques"** being submitted by Mr. **Dinesh Kumar Vishwakarma** (Reg. No.: 2K11/PHD/EC/11) for the award of degree of Doctor of Philosophy to the Delhi Technological University is based on the original research work carried out by him. He has worked under my supervision and has fulfilled the requirements 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 has neither partially nor fully submitted to any other university or institution for the award of any degree or diploma.

Prof. Rajiv Kapoor
(Supervisor)
Department of Electronics & Communication Engineering
Delhi Technological University

ACKNOWLEDGMENTS

I owe my debt and would like to express deep feelings of gratitude to accomplish the research program with the support and direction of several persons. This challenging and rewarding experience definitely helped me to grow in character as well as academically. It is a pleasant aspect that I have now the opportunity to express my gratitude for all of them.

First and foremost I would like to thank my supervisor **Prof. Rajiv Kapoor**, for his support and trust throughout the course of my Ph.D. studies. His substantial and thorough approach, together with his genuine interest in the research subject, turned my research work into a great experience. Words cannot express my gratitude to him for his patience and support.

Above all, my deepest thanks go to my parents for their unconditional love and support. Also, my heartiest thanks go to my family, my daughter **(Avika)**, my wife **(Sushma)** whose proximity, love & affection, and whose everyday prayers made it possible to complete this research work. I admire their sincere efforts in providing support at various stages of the Ph.D., otherwise, it would have been impossible to finish the long journey of the Ph.D.

I would also like to express my sincere thanks to my post graduate (Master of Technology) supervisor **Prof. I.K. Bhat**, for the encouragement and advice, he has provided during the post-graduation as his student.

Finally, I would like to express my thanks to my friend **Mr. Kuldeep Singh**, my faculty colleagues, my students and all those who supported me directly or indirectly.

Date: (**Dinesh Kumar Vishwakarma)**

Place: Delhi

ABSTRACT

With the advent of technology and increasing demand of society, video sequence analysis based systems are becoming the reality of various applications and active research areas of the computer vision. The wide range of applications include video based intelligent surveillance, motion analysis, video indexing, web based video filters, human-computer interaction, human activity recognition, object tracking, smart interactive televisions, animations and special effects in movies, sports analysis and so forth. The main building blocks of a video sequence analysis system consist of pre-processing, features extraction and representation, and classification.

In view of the various applications of video sequence analysis, this thesis investigates human activity recognition approach based on human silhouettes. Human silhouette is the basic information unit for representation and recognition of human activities. To achieve higher recognition accuracy of human activities, a three step methodology is devised:

- First step is extraction of human silhouette using texture based foreground segmentation;
- The second step is extraction and representation of features, which is done using two major approaches: first is the grids and cells based and the second is computation of spatial distribution of gradients and rotation of human silhouette;
- The third step is classification of human activities performed by various state-of-the-art classifiers.

This research is mainly focused on proposing novel approaches for extraction and representation of features in the action and activity recognition methodology based on human silhouette.

In the first approach, the key poses of the human silhouettes are selected based on the high energy principle. These key poses are divided into various cells and further features are computed. The computed features are then represented in such a manner that the spatio temporal information of the human silhouette is maintained. The represented features are used to form a feature vector and these feature vectors are classified using linear discriminant analysis (LDA), K-nearest neighbour (K-NN), and support vector machine (SVM). The recognition accuracy achieved on various publicly available dataset is compared with similar state-of-the-art methods. The proposed approach has demonstrated superior performance.

The key ingredients of second approach are the spatial distribution and rotation of human silhouette which include: the formation of average energy silhouette images; computation of magnitude and gradients of the pixels; sum of direction pixels and rotation of human silhouette. The spatial distribution of average energy images is computed by determining the magnitude and gradients of each pixel and further these magnitudes are quantized into orientation bins. The sum of directional pixels is computed by summing the pixels values in x and y directions. The rotation of binary human silhouette is computed using $\Re$-transform. Based on these features, a feature vector is formed by concatenation, which results a novel descriptor for the recognition of human action and activity. The performance of formed descriptor is tested with various publicly available datasets and compared with earlier state-of-the-art algorithms.

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

LIST OF PUBLICATIONS

- **D.K. Vishwakarma,** Rajiv Kapoor, “Hybrid classifier based human activity recognition using the silhouette and cells”, ***Expert Systems with Applications***, Vol. 42, No. 20, pp. 6957–6965, 2015. (**Publisher: Elsevier**)
- **D.K. Vishwakarma,** Rajiv Kapoor, "Integrated Approach for Human Action Recognition using Edge Spatial Distribution, Direction Pixel, and R - Transform", ***Advanced Robotics***, Vol. 29, No. 23, pp. 1551-1561, 2015. (**Publisher: Taylor & Francis**).
- **D.K. Vishwakarma,** Rajiv Kapoor, "An Efficient Interpretation of Hand Gestures to Control Smart Interactive Television" ***International Journal of Computational Vision and Robotics (IJCVR),*** http://www.inderscience.com/info/ingeneral/forthcoming.php?jcode=ijcvr (***In Press)*** (July, 2015). (**Publisher: Inderscience, UK**).
- **D. K. Vishwakarma** et al., “A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics,” ***Robotics and Autonomous Systems***, vol. 77, pp. 25-38, 2015. (**Publisher: Elsevier**)
- **D. K. Vishwakarma** et al., “Unified framework for human activity recognition: An approach using spatial edge distribution and $\mathfrak{R}$-transform,” ***AEU - International Journal of Electronics and Communications,*** vol. 70, no. 3, pp. 341-353, 2016. (**Publisher: Elsevier**).
- **D.K. Vishwakarma** et al., “Human Activity Recognition Using Gabor Wavelet Transform and Ridgelet Transform”, *Procedia of Computer Science Journal Vol. 57C,* pp. 630-636, (2015). **(Publisher: Elsevier*).***

- **D.K. Vishwakarma** et al., “Human Motion Analysis by Fusion of Silhouette Orientation and Shape Features” *Procedia Computer Science Journal, Vol.* 57, pp. 438-447, 2015. **(Publisher*:* Elsevier*).***
- **D. K. Vishwakarma** et al., “Recognition of Abnormal Human Activity using the changes in orientation of silhouette in key frames” *Proc. of the 9th INDIACom-2015, 2nd IEEE International Conference on “Computing for Sustainable Global Development”,* New Delhi, India, 2015.
- **D.K. Vishwakarma** et al., “A Novel Approach for the Recognition of Hand Gestures from Very Low Resolution Images”, *Proc. of 6TH IEEE international conference on communication system and network technologies*, Gwalior, India, 2015.
- **D.K.Vishwakarma,** Rajiv Kapoor “Simple and Intelligent system to recognize the expression of speech disabled person” *4th IEEE international conference on Intelligent Human Computer Interaction,* pp: 1-6, Kharagpur, India, 2012.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION TO VIDEO SEQUENCE ANALYSIS

This chapter introduces first the background of video sequence analysis, which includes the need of video sequence analysis, applications, and difficulties involved therein. Thereafter, we describe one of the important applications of video sequence analysis that is human activity recognition which includes the systematic flow diagram of human activity recognition system, various steps used like segmentation, feature extraction and representation, and classification. The challenges involved in each step are also discussed. At last, the significance of study and overview of the thesis is explained.

1.1 Video Sequence Analysis

A sequence of images displayed at a certain rate of frequency (frames per second) is termed as a video sequence. It consists of information in the form of spatial changes with respect to time, and if someone wants to extract the information from a video sequence then the spatial change with respect to time must be perceived. For the video sequence analysis (VSA), the knowledge of core technologies of digital image processing is a fundamental requirement, which includes image enhancement, image segmentation, morphological operations, feature extraction and representation, image classification etc. The main aim of this analysis is to automatically detect and determine the spatio-temporal events in the video signal and to further interpret the nature of event. In this analysis, the identification of appropriate objects/regions in the scene (segmentation), and the most descriptive characteristics of each object/region in the complete scene (feature extraction) are extracted. In both the cases, segmentation and feature extraction, the spatial and temporal dimensions must be taken into consideration for effective representation of objects/regions. Feature extraction is used for both coding and indexing and the most adequate coding parameters for each of the objects/regions is set. A similar process is used for segmentation, to identify the objects/regions, which permits particular and distinct coding and unwraps enhanced interaction possibilities. Similarly, the capability to define content in an object/region-based technique increases the accuracy of the description.

Since last few decades, VSA using intelligent techniques has emerged as a promising and interesting area of research in the field of computer vision and image processing due to the critical issues and numerous applications [1] [2] [3] of this analysis.

1.2 Critical Issues in Video Sequence Analysis

In VSA, there are numerous factors which limit the performance of VSA system such as recording settings, illumination variations, camera motion, view point variations, background complexity, similarity between foreground and background objects, high dimensionality and redundancy of the data etc [4]. All these factors provide an open challenge to the researchers/technocrats, to design and develop such an algorithm which has the capability to deal with these issues.

The environmental conditions play a very important role when the recording/ acquisition of the video signal is done because the performance of vision based system is highly dependent on the weather conditions. The captured video signal in bad environmental conditions leads to a poor quality of video signal. Due to the poor quality of video signal, the object and background of the scene may not be discernible and because of this, the subsequent task (segmentation, feature extraction) associated with video sequence analysis system leads to worse performance [5]. Hence, proper illumination is needed to acquire good quality video signal, where object and background are discernible.

The motion of the camera causes blurred image of the object in the scene and due to this an additional de-blurring algorithm is needed to de-blur the object [6]. Hence, to avoid this there must be proper installation of the camera.

The complexity of the background introduces the problem of extracting the object from the scene. Due to cluttered background, the object (foreground) and background may have more similarity [7] and because of this, the accurate segmentation of the object may not be possible. Hence, it is worthwhile to note that the recording of the scene must be performed based upon the application to avoid these issues.

1.3 Applications of Video Sequence Analysis

There are numerous applications of VSA and these are broadly categorised as:

- **Entertainment:** one of the important applications which is directly related to the daily life of human beings is entertainment, in the form of television (TV), movies, high definition television (HDTV) transmission, video games, live streaming of sports analysis etc [8] [9] [10].

- **Commercial:** VSA can be used for commercial purposes such as smart CCTV, advertisement of product, and also in retail industry for tracking the shoppers inside the store etc [11].

- **Security and Surveillance:** Most widely used application for security and safety purposes by military and police [12]. It can be for instance, monitoring of crowd behaviour at public functions, detecting/ preventing terrorist activities [13] at public places like airports, railways stations, bus stands etc., robbery detection, home intrusion system etc.

- **Human computer interaction:** Nowadays the interaction of human being with machine is increasing rapidly [14] due the advancement of VSA technologies. The traditional way of interaction of humans with machines is through remote control, keyboard, mouse, joystick etc. but in the coming years these modes of interactions may become obsolete due to the invention of various recognition systems based on [15] [16] body postures, hand gestures, facial expressions, etc.

- **Motion Analysis:** There are a variety of systems, where VSA is used to detect and determine the motion of object. The effective detection, tracking, and recognition of an object leads to several important applications like human activity recognition system [17] for detecting various kind of human abnormal and normal activities, object tracking, intruder detection and industrial monitoring.

As highlighted, VSA has a number of applications in various fields of science and technology but the main focus of this research is to design and develop a novel VSA algorithm for human activity recognition (HAR) system.

1.4 Human Activity Recognition System

A vision based human activity recognition system is capable of automatically detecting and determining the ongoing activity in the video sequence by extracting and interpreting the spatio-temporal changes in the video sequence. As it is depicted in Figure 1, the HAR systems in general have some key processing steps like pre-processing, feature extraction and representation, and classification or recognition of an activity.

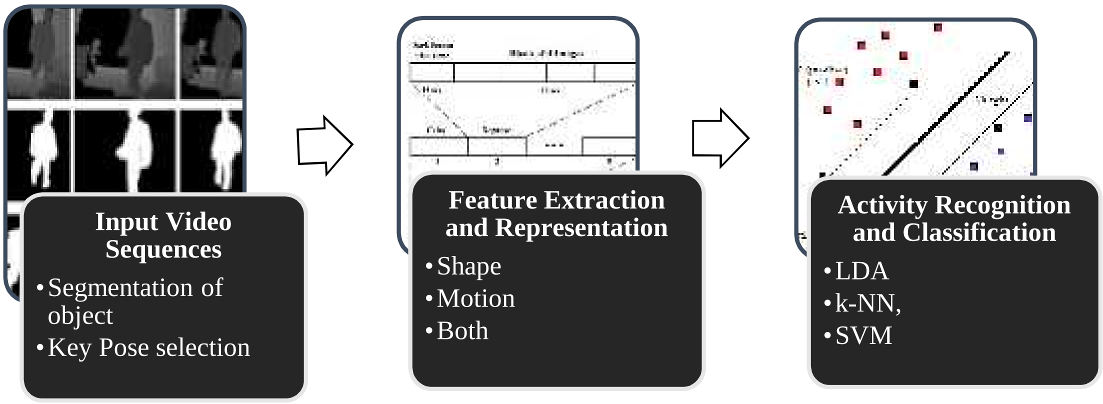


Figure 1.1: Overview of HAR system.

In recent years, the area of vision based Human Activity Recognition (HAR) has grown phenomenally, reflecting its importance in many high impact social applications including intelligent surveillance, web video search and retrieval, elderly care system for better quality of life, content based video analysis, interaction between people, sports analysis, intelligent robotics, and prevention of terrorist activities [1] [18] [19] [20] [21]. The typical task of a HAR system is to detect and analyse human activity in a video sequence. The reviews of previous work [1] [3] [22] [20] [23] reveal the challenges in vision based HAR systems. Various factors that make the task challenging are the variations in body postures, the rate of performance, lighting conditions, occlusion, view point and cluttered background. A good HAR system is capable of adapting to these variations and efficiently recognizing the human activity class. The important steps [24] involved in HAR systems are usually: a) Segmentation of foreground b) Efficient extraction and representation of feature vectors, and c) Classification or recognition. An effective and novel solution can be proposed at any step of the work individually, or collectively for all the steps. Due to the variations in human body anatomy and environmental conditions, every step is full of challenges and therefore, one can only provide the best solution in terms of recognition accuracy and processing speed. The shape and motion feature based descriptors [1] are two widely used methods in HAR systems. Shape based descriptors are generally represented by the silhouette of the human body and silhouettes are the heart of the activity. Motion based descriptors are based on the motion of the body, and the region of interest can be extracted using optical flow and pixel wise oriented difference between the subsequent frames. The motion based descriptors are not efficient, especially when the object in the scene is moving with a variable speed. The subsequent section highlights the detail of issues related to HAR system and its functioning.

1.4.1 Major Challenges of Human Activity Recognition System

The major challenges of vision based HAR system are to deal with cluttered background, motion of the camera, view point variations, illumination change, occlusion, intra-class dissimilarity and inter-class similarity etc. The cluttered background creates a difficulty for selection of foreground objects (segmentation) because of the disorderly arrangement of the objects in background and the object present in the background and foreground may have similar characteristics. For human silhouette segmentation, the prime step is to extract the human silhouette accurately for effective representation of human activity. Due to sudden motion of the camera and view point change, the captured image is distorted and therefore, it is vital that the position of camera should be fixed otherwise some additional process are required to detect the motion of the camera and view point change [25], which may be a complex task. The performance of vision based system is highly affected by the lighting conditions. Due to poor lighting condition, the captured video signal has low intensity pixel variation and extraction of desired object becomes a difficult task. Hence, it becomes, a necessity to maintain proper illumination, throughout the day but it isn’t always practically possible, especially in the case of surveillance applications. Therefore, again it becomes necessary to maintain proper illumination, where the HAR system is installed, but up to a certain level of pixel intensity, variation can be adjusted by using proper enhancement techniques [26]. Due to the occlusion of the object, the complete information about the object cannot be captured by single camera and the reason for occlusion in HAR system may be due to self-body part or alignment with another person. The effect of occlusion can be restricted in HAR system by incorporating multiple cameras, which give information at multiple views [27], or by employing a robust technique [28] for feature extraction but these solutions have their own limitations. Having multiple cameras, increases the cost and complexity of system while robust feature exaction technique may be suitable only to track single person activity. Inter-class similarity and intra-class dissimilarity of different kinds of human activities plays a crucial role for the classification. To improve the classification accuracy of the HAR system, a robust classifier which has the capability to deal with the high interclass similarity and intra-class dissimilarity, is required [29].

1.5 Problem Statement

Video sequences of different human activities are given, which have clothing variation, zoom in, zoom out, illumination variation, high intra-class dissimilarity and inter-class similarity. Under these circumstances, a framework of solutions is proposed that detect and determine the human silhouette of the activity, and based upon the human silhouette configuration, human activity is recognized and classified. A texture based entropy model provides the solution of accurate silhouette extraction under such variations. The problem of redundancy, losing geometrical and temporal information of feature representations are solved through key pose selection, cells and grid, computation of spatial distribution and the sum of directional pixels, and □-transformation. In order to address the problem of classifier to deal with interclass similarity and intra-class dissimilarity, a robust multi-classifier is designed using the fundamentals of support vector machine.

1.6 Main Contribution of the Thesis

The main contribution of this thesis is to design and develop various novel approaches for improving the recognition accuracy of HAR system. This thesis also gives the theoretical basis for the improvement of performance of HAR methods.

1.6.1 Theoretical Formulation

- The problems of segmentation, and spatial and temporal redundancy of the video signal have been identified.

- The issues involved in the recognition of human activity under the various lighting conditions have been featured.

- An issue related to the representation of 3-dimensional video signals has been identified and dealt with.

- Computation of the motion based temporal information of moving human beings is exhibited.

- The performance of the classifiers under various constraints of the activity performed has been observed.

- The low recognition accuracy of the HAR system under complex human activity has been detected.

1.6.2 Experimental Validation

- The issue of segmentation of object in a video sequence has been addressed by using a texture based entropy model.
- The spatial and temporal redundancy have been reduced by selecting key poses of human silhouette for an activity using highest energy concept.
- The representation of 3-dimensional video signal is done via 2-dimesional approach by forming average energy silhouette images (AESI).
- The loss of motion temporal information in AESI is compensated by computation of orientation of silhouette using □-transform.
- The effects of computation of spatial distribution of gradients at various levels, are computed and validated using standard activity datasets.
- The robustness of the proposed algorithms are validated using standard datasets.

1.7 Motivations of Human Activity Recognition

The main motivation behind the study of HAR system is its huge applications in real world and critical issues. Human activity recognition is a multidisciplinary area of research which is associated with neural networks, machine learning, intelligent computing, human computer interaction, as well as psychology and sociology. Thus, this field is drawing the attention of researchers for a variety of applications. Another important fact behind this work is to discover a novel framework for HAR system, which gives high recognition accuracy and in literature it has been observed that multiple feature [30] [31] [32] [33] based HAR system gives an improved performance in comparison with single feature based system. Hence, in this work also, some multiple feature based approaches for the HAR, are proposed.

1.8 Significance of the Study

The core finding of this study leads to opening of a wide framework for many real life applications, which are entirely based on human activity recognition. Recently, it has been seen that the HAR system demands are increasing day by day due their potential real world applications like unconstrained video search, aerial video analysis, sport video analysis, health care system, gait analysis and biometric recognition, intelligent robots etc. Another significance is the direction of future research, which may ignite the research community for further studies in this area with the help of current state-of-the-art.

1.9 Thesis Overview

In Chapter 2, the details of the earlier state-of-the-art methods, which include their merits and demerits in terms of pre-processing, feature extraction and representation, and classification are described. It also gives the highlights of the research gap in the concerned area, and based upon the research gap, the objectives of research are formulated and explained.

The silhouette and cells based bags-of-word model of activity recognition is explained in Chapter 3, which includes the detailed description of pre-processing (segmentation), feature extraction and representation, classification, experimental setup and discussion of results.

In chapter 4, activity recognition based upon the computation of spatial distribution of gradients, sum of direction pixels variations on average energy silhouette images, and human silhouette orientation based motion information computed using □-transform are presented. The detailed analysis of the computation of spatial distribution of gradients (SDGs) on average energy silhouette images at different decomposition levels is presented, and further, an effective model of SDGs computation is proposed, which is experimentally verified and validated on standard human activity datasets.

Chapter 5 highlights the important conclusions drawn from the research, and also gives the details of future scope of work.

CHAPTER 2

LITERATURE REVIEW

This chapter introduces the details of earlier work carried out for each and every step of HAR system, which include the pre-processing approaches to select the human silhouette, types of feature extraction and representation methods, the methods used for dealing with high dimensionality of the feature vector and various classification models used for the recognition and classification of human activities.

2.1 Pre-processing

The video signal captured by the digital camera/CCTV, in most scenarios, is required to be pre-processed for the later steps. The pre-processing steps are generally enhancement, conversion of images (RGB to Gray), segmentation, selection of region of interest (ROI) and normalization. The detection of human in a video sequence is an extremely difficult task due to the illumination changes and camera movement. The main objective is to reduce the complexity of the task by considering all the required information, then building a model or an algorithm that makes the human detection task simple and reliable. In this case, the most important information is that the human present in the video sequence is moving and the goal is to extract the human present in the video. The extraction can be done by considering the background to be static, periodically moving or dynamically moving. At pre-processing stage, the main focus is to reduce the region of interest by using background subtraction (BS) or optical flow based method of segmentation. A variety of algorithms are used for these methods and some popular methods are discussed here.

2.1.1 Background Subtraction

The BS is used where the human is moving and camera is stationary. The simplest way to get the background image is to capture a background image, which does not have the moving human in the scene. For each video frame, the absolute difference between the current frame and static image is computed which is termed as static frame differencing. However, this method of foreground detection may fail when the illumination is changed. In most of the realistic situations, where the background image is not available then a background modelling approach is used to construct the background image. There are many approaches used for background modelling [34] [35] [36] [37]. These methods mostly differ in the way backgrounds are modelled and are broadly classified as [38]: Basic model, Statistical model, Fuzzy model, and other models. Some of the fundamental background modelling approaches which are frequently used for the segmentation of foreground objects are classified as: fundamental approaches and statistical approaches.

2.1.1.1 Basic Modelling

In basic modelling approach, the immediate previous frame is considered as background frame and it works only for particular conditions of the object i.e. speed and frame rate. It is also very sensitive to the threshold and mathematically it can be defined for a video sequence $f(x,y,t)$ with length ℓ, which is the total number of frames of the video as:

$$\beta(x,y,t) = f(x,y,t-1) \tag{2.1}$$

$$F_G(x,y,t) = f(x,y,t) - \beta(x,y,t-1) > Threshold \tag{2.2}$$

where Equation 2.1 represents the background image $\beta(x,y,t)$ and Equation 2.2 is the foreground image $F_G(x,y,t)$, which consists only of the object. This method of extracting the object is simply known as frame differencing approach and working of this approach is highly dependent upon the object structure, speed, frame rate, and threshold. Due to these constraints, this method of getting the object is not used for human detection, especially when the human is performing an activity in the video sequence. This approach also fails when the object in the scene suddenly stops. Therefore, to address the problem of this method, a modelling approach suggested by Lai and Yung [39] to update and initialize the background model by taking the running average of the pixels of successive frames in the video sequence, is considered. The averaged image is denoted as background image and it is modelled as:

$$\beta(x,y,t) = \frac{1}{n}\sum_{i=0}^{n-1} f(x,y,t-i) \tag{2.3}$$

Equation 2.3 gives the mean image of the previous n frames and this is considered as the background frame but the accuracy of this modelling depends on the object speed and frame rate, and also requires high memory. To make this modelling more robust, a mechanism to update and maintain is incorporated by recursively computing the background image at each instant of time, which is expressed as:

$$\mathscr{b}(x,y,t) = \frac{t-1}{t}\mathscr{b}(x,y,t-1) + \frac{1}{t}f(x,y,t) \tag{2.4}$$

Equation 2.4 can be generalized and represented by Equation 2.5:

$$\mathcal{B}_t = (1-\tau)\mathcal{B}_{t-1} + \tau F_t \tag{2.5}$$

where $\mathcal{B}_t$ is the background model at time t and is related as $t \in \{1, n\}$, τ is the learning rate and is related as $\tau \in [0,1]$, and F_t is the video sequence. The key advantage of this method is that it automatically updates the background model but it may fail under a few circumstances such as when the background is bimodal, and the moving human is fast and frame rate is slow.

Recently, some methods [40] [41] [42] [43] [44] suggested the use of colour, texture and edges of the scene to perform foreground detection. Chua et al. [40] proposed a robust colour texture based BS model with an adaptive weighting scheme that automatically modifies the weight between the colour and texture similarities. Zhang and Xu [41] introduces a novel colour and texture feature based fusion model for background subtarction. A hierarchical coarse-to-fine texture description based background modelling approach presented by Yeh et al. [42] , fully utilizes the texture characteristics of each preceding frame and the method can handle both the shadow disturbance and lighting variation. Chiranjeevi and Sengupta [43] proposed a new algorithm for the detection of moving object using combination of intensity and statistical texture features for better object localization and robustness. Martínez et al. [44] proposed a method to assess the texture feature based on coarseness, contrast and directionality. These features play a vital role in the human perception of texture. To meet the purpose, they used the fuzzy set of different groups of measures and the performance of each set is analyzed with human assessment.

2.1.1.2 Statistical Modelling

In the statistical approach of BS, each pixel histogram is modelled by a single Gaussian function [45] or a mixture of Gaussian functions [46] or kernel density estimation [47] and the updating is done by running average. The foreground and background classification of the pixels is done using statistical variables i.e. mean and variance. In these models, if a pixel is classified as a background pixel then it is ignored in the foreground. Wren et al. [45] describe that a single Gaussian model is used for each pixel histogram but it doesn’t handle the multimodal backgrounds. To address the multimodal background problem, a model is proposed in [46] , which is called as Gaussian mixture model (GMM) but the issues with GMM are handling of number of Gaussians, initialization parameters and updating over the time. In GMM, each pixel colour distribution is represented by weighted Gaussian distributions in a certain colour space. The distributions are updated by running average by computing mean (μ) and variance (σ^2) through Equation 2.6 and 2.7, which are as follows:

$$\mu_{t+1} = \tau F_t + (1 - \tau)\mu_t \tag{2.6}$$

$$\sigma^2{}_{t+1} = \tau(F_t - \mu_t)^2 + (1 - \tau)\sigma^2{}_t \tag{2.7}$$

A new image is processed by GMM (m_t) at time t from the measures ($f_0, f_1, \dots . f_{t-1}$) of each correct pixel. The probability that a pixel is a part of background or foreground is defined as:

$$P(f_t/m_t) = \sum_{n=1}^{N} \frac{\tau_n}{(2\pi)^{d/2}|\Sigma_n|^{1/2}} \exp(-\frac{1}{2}(f_t - \mu_n)^T \Sigma^{-1}(f_t - \mu_n) \tag{2.8}$$

where f_t is the measure and d is the dimension of colour space, N is the number of Gaussians, ∑ is the covariance matrix, τ_n is the weighting factor, where $\sum_{n=1}^{N} \tau_n = 1$. |.| is matrix determinant. In RGB colour space if each pixel is considered as independent then covariance matrix is defined as:

$$\sum_n = \begin{pmatrix} \sigma_{1,n}^2 & 0 & 0 \\ 0 & \sigma_{2,n}^2 & 0 \\ 0 & 0 & \sigma_{3,n}^2 \end{pmatrix} \tag{2.9}$$

where the subscripts 1, 2, 3 refer to channel numbers like the RGB colour space. To update the GMM, first f_t is measured, which is associated with one Gaussian out of n, if:

$$\| f_t - \mu_n \| < k\sigma_n \tag{2.10}$$

where k is 2 or 3 [38], σ_n represents the variance of Gaussian distribution of index n. If the condition in Equation 2.10 is true then this measure represents the background of the image. The performance of GMM is good in outdoor scenes and it can also handle a little variation in the lighting conditions. Hence, it may be a good BS algorithm for video surveillance applications but the shadow of the objects creates a problems and it is also inefficient, if the video frames are noisy.

2.1.2 Optical Flow

Optical flow is defined as the flow of intensity pattern or displacement of pixels in subsequent frames of a video sequence. For the extraction of moving object in a video sequence, a variety of methods [48] [49] are used to estimate the optical flow field in the subsequent frames. Optical flow reflects the change in an image due to motion in consecutive frames. It also reflects that the three-dimensional motion of an object points across a two-dimensional image and it can state about the relative distances of equal speed objects. There are number of approaches to compute the optical flow by different ways but mostly they have three common stages of processing [50], which are as follows:

- In order to extract signal with enhanced signal to noise ratio, pre-filtering and smoothing is performed.

- Secondly, the basic measurements such as spatio-temporal derivatives and local correlation extraction are done.

- In the last stage, integration of these measurements is done to produce 2D- flow field, which often includes assumptions about the smoothness of the original flow field.

From the recent reviews [51] [52], optical flow computation approaches can be broadly categorised based on derivative, region, frequency and phase of an image.

The fundamental approach is based on the derivative of an image, in which optical flow of image frames is computed using the spatio-temporal derivative by considering few constraints like brightness constancy, velocity smoothness, and temporal persistence. To understand better, let us assume an image frame, where $I(x,y,t)$ is the center pixel of $M \times N$ image and this pixel moves with $\delta x, \delta y$ in time δt and denoted as $I(x+\delta x, y+\delta y, t+\delta t)$. As it is assumed that intensity is constant hence it can be written as:

$$I(x,y,t) = I(x+\delta x, y+\delta y, t+\delta t) \tag{2.11}$$

As it is assumed that $\delta x, \delta y$ and δt are partial changes of the location with time, hence Equation 2.11, can be expressed using Taylor series expansion as:

$$I(x+\delta x, y+\delta y, t+\delta t) = I(x,y,t) + \frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t + h.o.t. \tag{2.12}$$

where h.o.t. stands for higher order terms and it is assumed that these are very small hence can be omitted. Using Equation 2.11 and 2.12, it can be expressed as:

$$\frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t = 0 \text{ or } \quad \frac{\partial I}{\partial x}\frac{\delta x}{\delta t} + \frac{\partial I}{\partial y}\frac{\delta y}{\delta t} + \frac{\partial I}{\partial t} = 0 \tag{2.13}$$

$\frac{\delta x}{\delta t}$ and $\frac{\delta y}{\delta t}$ is the displacement upon time, which can be called as image velocity or flow of intensity, or optical flow and respectively written as: v_x and v_y . The partial derivative in Equation 2.13 is normally written as:

$$\frac{\partial I}{\partial x} = I_x, \frac{\partial I}{\partial y} = I_y \text{ and } \frac{\partial I}{\partial t} = I_t \tag{2.14}$$

Equation 2.14, can be compactly written as:

$$(I_x, I_y).(v_x, v_y) = -I_t \text{ or } \square I.\vec{v} = -I_t \qquad (2.15)$$

where □I = (v_x, v_y) is the spatial intensity gradient and $\vec{v}$ = (v_x, v_y) is the optical flow at pixel (x, y) at time t. Equation 2.15 is also called fundamental optical flow computation equation. The fundamental principle of computation of optical flow was explained but later on some modifications were proposed by [53] [54] to make optical flow techniques more robust. There are certain issues with the optical flow methods like the problem of aperture, need of uniform illumination, shadows of objects, and occlusion. Using this principle of computation (optical flow), the moving pixels locations can be computed and based upon the need, further tasks can be performed on these pixels like segmentation of moving objects, motion feature computation, object tracking etc. After the pre-processing stage, the feature extraction and representation are carried out in HAR system.

2.2 Feature Extraction and Representation

In vision based activity recognition, an activity is recorded or captured by the video camera/CCTV. Activity is performed by the motion of the whole human body or some part of the human body. Due to the motion of human body/part the shape of human silhouette changes with respect to time. Hence, it can be said that for the recognition of human action and activity, one must extract shape as well as the motion of human body in the consecutive frames of the video sequence.

In the past, a significant amount of work has been reported in the literature for the recognition of human activity using video sequences and most of the HAR methods rely on the local features, global features, key points, spatial-temporal features, bags of words etc. [1] [3] [19] [20] [21] [55]. All these methods generate a set of features and then an appropriate machine learning classifier is used for the recognition of the activity.

In general, the HAR approaches of feature extraction can be categorized on the basis of feature representation and these are: learned geometrical approaches of human body parts, spatio-temporal templates, appearance or region features, shape features, interest-point-based representations, and optical flow or motion patterns. The details about these representations are formed and explained in the subsequent sections.

2.2.1 Visual Appearance based Representations

In recent years, appearance based representation of features has become popular. In this representation, the human body or parts of the human body are learned by appropriate models and further it is matched with the target video sequence for the activity recognition [31] [56] [57] [58]. The temporal information of the activity is computed by training of Hidden Markov Models (HMM) and their different alternatives. The appearance based approach works well for still image based activity recognition, where maximum visual information is present. However, activities involving the entire body are difficult to handle due to variation in clothing from one actor to other. A bag-of-word model based on textural appearance of humans is proposed in [58] and contains the formation of a new descriptor of space–time interest points that combines the descriptor of a 3D gradient extractor with a textural descriptor. Recently, Zhang et al.

[31], presented an appearance based anomaly detector using support vector data description.

2.2.2 3D-Volume based Representation

The approaches based on 3D volume of video for HAR use the features’ volumes, trajectories, and local interest points. In this approach, first the video image is constructed and then matched with the stored known representation. This approach of representation is very effective where background is stationary and object is moving. Also, it can construct 3D model of a human and represent the shape and motion spatially using motion energy images (MEI). By this approach the simple activities like sitting, waving and crouching can be effectively represented. Shechtman and Irani [59] presented a space time volume based approach, where background segmentation is not needed and a behavior-based similarity measure is used to determine whether two different space-time intensity patterns of two different video segments have similar underlying motion field. It detects the similar activity in the video sequence despite the differences in appearing pattern due to different clothing, different backgrounds, and different illumination. The benefits of this approach are that no prior modeling or learning are needed, and no need to build a complex model of human body configurations and the recognition can be accomplished directly on the raw video.

2.2.3 Interest Point based Representation

An efficient approach of spatio-temporal interest points (STIPs) based on local features using a temporal Gabor filter and a spatial Gaussian filter was introduced by Dollar et al. [60]. Thereafter, a number of STIPs detectors and descriptors have been proposed by several researchers [61] [62] [63] [64] [65] [66]. These local features based descriptors became popular due to their robustness against noise and occlusion because there was no need to track whole body of human being. However, these methods seemingly, are less effective for complex activity modelling (e.g. Ballet movement) and it is difficult to find the interest points.

2.2.4 Optical/Motion Flow based Representation

A motion template based activity recognition model has been presented by Hu and Boulgouris [67], where the templates are designed in such a way so as to keep the structural and motion information, which is the most discriminative among activities. The direct measurement of motion encapsulates the translational based motion information of the activity. In [68], optical flow of the scene is determined using Lucas-Kanade optical flow method with pyramid structure, which gives the motion information of the human to some extent. Then, this information is passed on to the feature descriptor, which consists of angles, bounce and other relevant information that discriminate the human activity. An approach [69] based on a set of kinematic features computed through optical flow method for HAR in the video sequence is presented and the computed features are divergence, vorticity, symmetry and antisymmetric optical flow fields, and flow gradients. Poularakis et al. [70] proposed a computationally efficient model for the recognition of human daily living activities based on the motion feature estimation. To compute motion feature a fast dense optical flow method is used, which is computationally efficient with minimum loss of recognition accuracy.

2.2.5 Shape and Silhouette based Representations

One of the most popular representation approach used for human activity recognition is shape based, because shape of an object is considered as a highly reliable parameter. This representation includes silhouette and edges of the human body but in general most of the work reported is based on the silhouette based representation. In this representation, human silhouette is considered as a fundamental information processing unit and this section is more relevant to the further research. Bobick & Davis [71] presented a silhouette based method in which the Motion History Images and Motion Energy Images (MHI, MEI) are used for activity recognition. These MEI and MHI are the images extracted from the video frames and these images are then stacked so as to preserve the temporal content of the activity. A method [72] based on contour points of human silhouettes is used to represent the human poses, and the actions are classified by using multi-view key poses. Later on, Chaaraoui et al. [73] proposed another method of HAR system based on silhouette, where they optimized the parameters using evolutionary computing and reported enhanced performance. Silhouette based approaches give effective representation of human activity but they result in high feature vector size and are less suitable when human is occluded. A holistic approach of human action recognition that relies on the human silhouette sequences was proposed by several researchers [22] [71] [73] [74] [75] [76] [77]. In silhouette based method, the foreground is extracted using background segmentation and then features are extracted from the silhouettes. Weinland et al. [22] worked on matching template technique in which the region of interest (ROI) is divided into a fixed spatial or temporal grid due to which, the effect of noise present in an image and viewpoint variance can be reduced significantly. Thurau & Hlavac [78] used a histogram of oriented gradients based approach to represent activity, and also concluded that silhouettes are the best information unit for representing human activity. Shao et al. [77] proposed modified bags-of- words model called as bag of correlated poses using the advantages of global and local features. They addressed the problem of losing geometric information in bag of visual words representation, which generally is implemented using k-mean clustering algorithm.

More recently, multiple feature based techniques [30] [31] [32] [33] [79] are becoming popular for the recognition of human activities. Sedai et al. [32] proposed a multiple feature based HAR model, which combines the shape and appearance features, which gives the improved performance over the single feature based methods for HAR. Shao et al. [80] use multiple features for the recognition of human activity and these feature are 3D gradients and wavelet transform.

In all these approaches, it has been observed that the holistic approach model results in high dimensionality of the descriptor; hence there is a need for dimensional reduction techniques for efficient recognition.

The PCA [81] is a popular linear dimensionality reduction technique that has been widely used for dimension reduction and classification purpose in activity recognition [75] [82]. The determined data of the feature set has the correlated and uncorrelated data with each other. The correlated feature set of different classes makes the classification complex, and slow. Hence, for efficient and fast classification, feature set must have uncorrelated set and the uncorrelated feature set has lower dimension.

Therefore, for easy handling of data and improved classification, the dimension of feature set must be reduced. The PCA is a popular method for the reduction of dimension of the feature set by maximizing the variance of the feature set and mapping the feature sets into a lesser dimensional space. The concept of PCA technique for dimension reduction and getting the uncorrelated feature set is as follows: Consider a feature matrix $\mathcal{M}$ having t-dimensions. The principal axes $\{\mathfrak{D}_1, \mathfrak{D}_2, \mathfrak{D}_3, \dots \dots, \mathfrak{D}_n\}$ with $1 \leq \quad \leq t$, are the orthogonal axes on which the variance is maximum in the project space. Generally, $\{\mathfrak{D}_1, \mathfrak{D}_2, \mathfrak{D}_3, \dots \dots, \mathfrak{D}_n\}$ are expressed by principal Eigenvectors of the sample covariance matrix as:

$$C = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} (x_i - \mu)^T (x_i - \mu) \tag{2.16}$$

where $x_i \in \mathcal{M}$, μ is the mean of the samples, $\mathcal{N}$ is the number of samples. The orthogonal directions and largest variance can be found from Eigenvectors, as expressed:

$$\mathcal{W}\mathfrak{D}_i = \lambda_i \mathfrak{D}_i, \qquad i \in 1, \dots \dots \tag{2.17}$$

where, λ_i is the i^{th} largest Eigen value of $\mathcal{W}$. The principal component of given observation vector of $x_i \in \mathcal{M}$ is expressed as:

$$p = [p_1, p_1, p_1, p_1 \dots \dots p_n] = [\mathfrak{D}_1^T x, \mathfrak{D}_2^T x, \mathfrak{D}_3^T x, \dots, \mathfrak{D}_n^T x] = \mathfrak{D}^T x. \tag{2.18}$$

where, p is the principal component of x. The reduced dimension of feature set leads to a better classification performance with high recognition accuracy and fast computation time and these effects can be observed by classifying the feature set with different classifiers.

2.3 Classification Models

There are various types of classifiers, which are used for the purpose of classification and recognition of human activities. Mostly for human activity classification, machine learned classifiers are widely preferred due to their intelligent capability of handling unknown test samples. Some of the most widely used classifiers are linear discriminant analysis (LDA), K-nearest neighbour (KNN) and support vector machines (SVM).

2.3.1 Linear Discriminant Analysis

It provides an improved classification of various classes of data by maximizing the margin between dissimilar classes and minimizing the margin within the same class. The scatter matrices within the class and between the classes can be determined as [83]:

$$\mathfrak{S}_{\mathcal{W}} = \frac{1}{N}\sum_{i=1}^{\mathfrak{C}} \sum_{x\in\mathfrak{C}_i} (\mu_i - \mu)(\mu_i - \mu)^T \tag{2.19}$$

$$\mathfrak{S}_{\mathfrak{B}} = \frac{1}{N}\sum_{i=1}^{\mathfrak{C}} n_i (\mu_i - \mu)(\mu_i - \mu)^T \tag{2.20}$$

here, *N* represents the total number of samples, $_i$ denotes the number of sample points of each class used for training, x is a sample vector of a specific class, $\mathfrak{C}$ is the total number of classes, μ_i is the mean of each class, μ is the total mean of the vector, $\mathfrak{S}_{\mathcal{W}}$ is the scatter matrix that represents the amount of scattering within the classes of activities and $\mathfrak{S}_{\mathfrak{B}}$ is the scatter matrix that signifies the amount of scattering matrix between the classes. To optimize the discrimination, the projection matrix in space can be determined as:

$$\varphi_{opt} = argmax \frac{\varphi^T \mathfrak{S}_{\mathfrak{B}} \varphi}{\varphi^T \mathfrak{S}_{W} \varphi} \tag{2.21}$$

where φ is the transformation and it can be expressed linearly as:

$$\mathfrak{S}_{\mathfrak{B}} \varphi = \lambda \mathfrak{S}_{W} \varphi \tag{2.22}$$

The maximum value of φ that corresponds to φ_{opt} can be obtained by obtaining the Eigen value λ. It has been proved that the largest value of λ corresponds to $\mathfrak{C} - 1$.

2.3.2 K-Nearest Neighbour

The K-nearest neighbour (K-NN) classifier selects the K-closest samples of training feature set to a new instance and the nearest class having highest votes is mapped to the test instance. One of the biggest advantages of this classifier is its non-parametric nature, it does not require any assumptions and easily classifies the data even in higher dimension space. Consider a query instance $\mathbb{Z}$, then the output of K-NN classifier is described as in [84]:

$$K - NN(\mathbb{Z}) = Max\ [p(\mathfrak{c}_i, \mathbb{Z})], \mathfrak{c}_i \in C \tag{2.23}$$

where $p(\mathfrak{c}_i, \mathbb{Z})$ is class probability and it is given as:

$$p(\mathfrak{c}_i, \mathbb{Z}) = \frac{\sum_{x \in K\mathbb{Z}} \ell(x_c = \mathfrak{c}_i).L(\mathfrak{d}(x, \mathbb{Z}))}{\sum_{x \in K\mathbb{Z}} K(\mathfrak{d}(x, \mathbb{Z}))} \tag{2.24}$$

where $\ell(.)$ is a function whose value is one, when the condition is 'yes', otherwise zero, $\mathfrak{d}(x, \mathbb{Z})$ is the distance between the two points which in this case is Euclidian distance and $K\mathbb{Z}$ is the nearest neighbour of $\mathbb{Z}$. The kernel function $K\big(\mathfrak{d}(x, \mathbb{Z})\big) = 1/\mathfrak{d}(x, \mathbb{Z})$.

The best matching label among the K points, which is closest to the training matrix, is chosen as the classification result.

2.3.3 Support Vector Machine

Support vector machine (SVM) is one of the most widely used classifiers for the classification of human activity [58] [85] [86] [87] due to high classification accuracy. The principle of SVM is based on structural risk minimization principle and the training is supervised. The label of the training data is used to obtain a decision boundary between the two classes. It is a non-probabilistic binary linear classifier and constructs a hyperplane in higher dimensional space which provides adequate separation between the two classes. The hyperplane is developed by maximizing the margin between the two classes. A margin is the distance between the two classes and is determined by the support vector points. The support vectors are the points that are nearest to the hyperplane. Maximizing the margin leads to the linear programming problem and due to this, it is also known as a maximum margin classifier. Support vectors are features of the samples, which are close to the hyperplane of an SVM, and are encircled in Figure 2.1 and 2.2. Therefore, the determination of the location of most important data is near the hyperplane and hyperplane is formed by the set of lines drawn between the class samples. To design an SVM, there are the two possibilities, one is that the classes are separable linearly and second is the classes are not separable linearly, which is as shown in Figure 2.1 and 2.2. To perform the classification of the two classes, a linear or nonlinear kernel based SVM is applied based upon the feature space. As it is seen from figure 2.1 and 2.2, where in Figure 2.1 the samples can be classified linearly by using linear function and in Figure 2.2 the samples cannot be classified linearly hence a kernel based on nonlinear classification approach may be used.

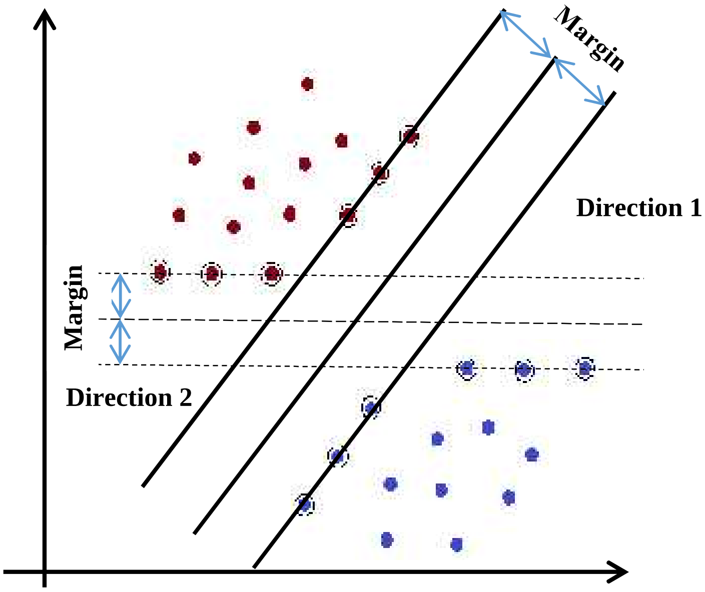


Figure 2.1: Two possible hyperplanes for a Linear SVM.

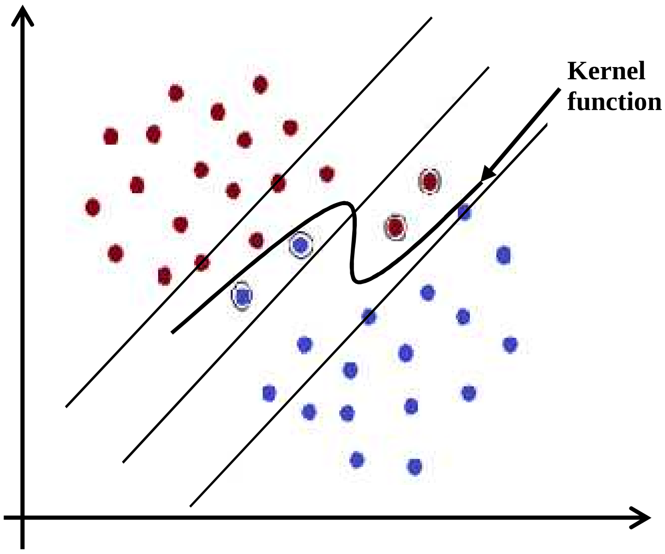


Figure 2.2: Non-linear function based SVM.

The classification in SVM is done by mapping the input data into the higher dimensional feature space. Consider *N* training samples of feature set $\mathcal{T}_k$. Where:

$$\mathcal{T}_k = \{(x_i, y_i) \parallel x_i \in \mathbb{R}^n , y_i \in (-1,1)\} \tag{2.25}$$

where *m* is the dimensional feature vector signifying the *i*^th^ training sample and $y_i \in (-1,1)$ is the class label of x_i. An optimal hyper plane can be expressed as:

$$\psi(x) = sgnf(\textstyle\sum_i^N \alpha_i y_i \mathcal{K}(x_i, x_j) + b) \tag{2.26}$$

Where $sgnf(.)$ represents the sign functions, and $\mathcal{K}(x_i, x_j)$ is nonlinear predefined kernel function. The non-linear SVM, which is formed by using radial basis function is as follows:

$$\mathcal{K}(x_i, x_j) = \exp(-p \parallel x_i - x_j \parallel^2 , \qquad p > 0 \tag{2.27}$$

The coefficients α_i and b can be determined using the concept of maximization of hyperplane value and are written as:

$$argmax\left[\textstyle\sum_i^N \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \mathcal{K}(x_i, x_j)\right] \tag{2.28}$$

where $\underbrace{min}_{\alpha} \quad \sum_i \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \ \forall i$ and $y_i = \square 1$. Where C is the penalty parameter that signifies the trade-off between maximizing the margin and minimizing the training set error.

Initially, it is used for classification of two classes, but later on it is extended to the multi-class problem. In two class problem, it is one of the most efficient and robust classifiers. Hence, by utilizing the advantages of the two class SVM, it is extended for multiclass SVM. The frequently used approaches of multiclass SVMs [87] are I) one-against-rest II) one-against-one III) directed acyclic graph SVM.

Although, the SVM is a binary classifier, it can still be extended for *M*-class classification for human activity recognition. One of the widely used multiclass techniques is one-against-rest [88] and the tool used for SVM classification is LIBSVM [89]. In one-against-rest classifier *M* number of binary SVM classifiers are formed, in which one class is separated from the rest of the classes. The i^{th} SVM is trained with all the training samples of i^{th} class with positive tags, and rest with negative tags. This process is done until all classes are trained against all others as one class. Testing of an unknown vector is done for all structures generated in the training and if this vector does not belong to any class then, it will lie in between the margin. And if this belongs to a class, then it will show up in that class, and in this way one can find to which class this vector belongs.

2.4 Research Gaps

Based upon the analysis of earlier state-of-the art methods on human action and activity recognition, we have captured the problems and listed a layout of the solutions for these problems, which are as follows:

- It is observed that the holistic representation of human activity requires an efficient method for the extraction of silhouette from the video sequence. Usually, foreground segmentation is done using background modelling and background subtraction, but it is not always possible to get good results due to inaccurate modelling of the background. Hence, in this work we have used a texture based segmentation approach to extract the human silhouettes.

- The problem of losing geometrical information in the bags of visual word is addressed by selecting key poses of the human silhouettes. Further, to describe the silhouette information, we have proposed a simple scheme which preserves the spatial change in the human silhouette over time.
- The loss of motion temporal information in average energy images (AEIs) are compensated by incorporation of additionally computed motion temporal information for action recognition.
- The performance of a classifier reduces when the activities have interclass similarity and intra-class dissimilarity. Therefore, to improve the classification of HAR system, we have constructed a hybrid classification model with the combination of “SVM-NN” classifiers.

2.5 Research Objectives

The main objective of this thesis is to analyse various issues involved in the human activity recognition using human silhouettes, and thus further to propose the solution framework based on the human silhouettes. Also, the problem of less classification accuracy of human action/activity is addressed by proposing a hybrid classification model. To fulfil these objectives, the following frame of work has been performed:

- For effective and accurate representation of human action and activity, an approach based on key poses of the human silhouette is developed by dividing the key poses of human silhouettes into cells and grids, which preserve the structural and temporal information.

- To address the problem of less recognition rate of human action/activity recognition system, a hybrid classification model is proposed, which is capable of dealing with the intra-class dissimilarity and interclass similarity.
- Design and development of a shape and motion rich feature descriptor by providing additional motion information to the average energy images computed through □-transform.
- The effect of computation of spatial distribution of gradients on average energy images at various levels is analysed. Further, based upon the analysis, a modified label of computation of spatial distribution of gradients is proposed, which gives the finer spatial distribution of shape.

CHAPTER 3

HUMAN SILHOUETTE REPRESENTATION

This chapter introduces the activity recognition using human silhouettes, which includes extraction of human silhouettes using texture based segmentation approach, feature extraction and representation using cells and grid approach, experimental setting, comparative analysis of result, and discussion of the result.

3.1 Introduction

The objective of this chapter is to present a new approach for human activity recognition in a video sequence by exploiting the key poses of the human silhouettes, and constructing a new classification model. The spatio-temporal shape variations of the human silhouettes are represented by dividing the key poses of the silhouettes into a fixed number of grids and cells, which leads to a noise free depiction. The computation of parameters of grids and cells leads to modelling of feature vectors. This computation of parameters of grids and cells is further arranged in such a manner so as to preserve the time sequence of the silhouettes. To classify these feature vectors, a hybrid classification model is proposed based upon the comparative study of Linear Discriminant Analysis (LDA), K-Nearest Neighbour (K-NN) and Support Vector Machine (SVM) Classifier. The effectiveness of the proposed approach of activity representation and classification model is tested over three public data sets i.e. Weizmann, KTH, and Ballet Movement. The comparative analysis shows that the proposed method is superior in terms of recognition accuracy to similar state-of-the-art methods.

3.2 Proposed Methodology

This approach is based on the silhouette of the human body which is extracted from the video sequence of the activity by segmentation techniques. The segmented silhouette is pre-processed to improve its quality for the feature extraction. Features generated from different silhouette images are then arranged in a representable form. Further, dimension reduction, and classification are used. The workflow diagram of the proposed framework is depicted in Figure 3.1.

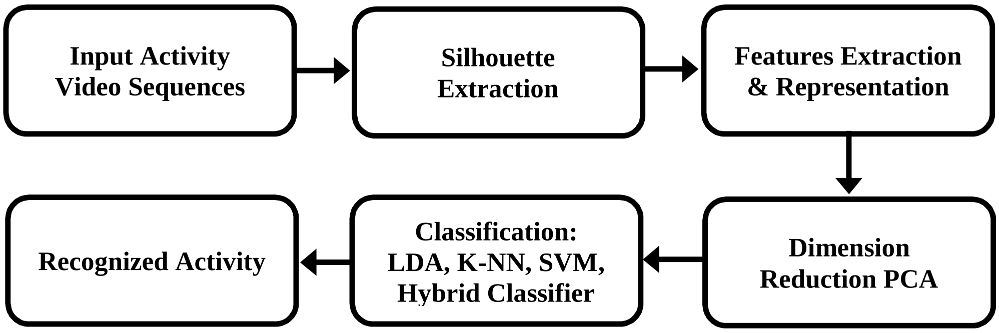

Input Activity Video Sequences
Silhouette Extraction
Features Extraction & Representation
Dimension Reduction PCA
Classification: LDA, K-NN, SVM, Hybrid Classifier
Recognized Activity

Figure 3.1: Workflow diagram of proposed framework.

Input of video sequence is provided by CCTV/Digital Camera and these sequences may be a sequence of colour images or grayscale images, depending upon the image pre-processing is carried out for later steps. The silhouette extraction is segmentation of foreground by elimination of background. The feature extraction and representation is crucial stage, which makes the algorithm good or bad. The represented features dimensions are generally high hence dimension reduction techniques is needed for fast classification and recognition of human activity and PCA is a popular techniques for this task. The detail description of each block is presented in following subsections.

3.2.1 Extraction of Human Silhouette

For human activity recognition, the background subtraction is considered as the fundamental challenge of vision based activity recognition. The other challenges which make the task challenging are illumination change, dynamic background, shadows, video noise, etc. [36]. In background subtraction, the fundamental concept is to construct and update the model of the background scene, and foreground object pixels are detected if they differ from the background model up to a certain limit. In the past, Gaussian Mixture Model (GMM) and Local Binary Pattern (LBP) based background models are widely used. In GMM, numerous Gaussian mixture distributions are used to demonstrate each pixel, where each Gaussian distribution characterizes the intensity of pixel distribution over time. For the noisy video sequence, the parameter estimation is unpredictable in case of assumption of Gaussian distribution. Hence, it can be concluded that this assumption is not always true. LBP is a very efficient textural operator, which labels the pixels of the image by thresholding the neighbourhood of each pixel, resulting in a binary pattern. Initially, Haralick et al. (1973), proposed a textural feature based segmentation approach using Gray Level co-occurrence matrix (GLCM). Thereafter, numerous textural feature based segmentation techniques have been proposed [40] [90] [91]. A robust textural feature based approach using LBP for background subtraction model is proposed by Chua et al. [40], in which, they have demonstrated that the textural feature based segmentation method gives more effective results for video surveillance applications. Textural feature based fast and efficient segmentation approach using Gray level co-occurrence matrix (GLCM) has been implemented on FPGA [90] [91] for different real life applications. The realization of texture based algorithm in real life applications has encouraged us to use textural feature based background subtraction in this work.

A method for describing different textures was originally presented by [92]. They proposed a matrix called Gray-Level Co-Occurrence Matrix, which allows describing texture based on differences in intensity in different directions and used 14 different features for classification of different textures in the image. Entropy is one of the most important parameter that describes the texture information in an image and it can be expressed as:

$$\zeta = \sum_i \sum_j \rho(i,j) \log(\rho(i,j)) \tag{3.1}$$

where $\rho(i,j) = \frac{M(i,j)}{\sum_{i,j} M(i,j)}$ is the probability density function; here i and j are indices to the co-occurrence matrix M. The entropy of the image is used to describe its complexity. Higher the entropy, higher is the complexity of the image. An entropy filter is generated in an image to represent the different textures present in the image. The filter matrix is generated for a pixel and its entropy is calculated in 9×9 neighbourhood. Converting this filter matrix into binary form with some thresholding gives an image with white spots at different areas. For a two-textured image, one part contains spots of the same size, and the same is true for the other parts. Removing one such part gives the mask for getting a human blob. Applying this mask over the raw image provides us with the silhouette from this raw image as shown in Figure 3.2.

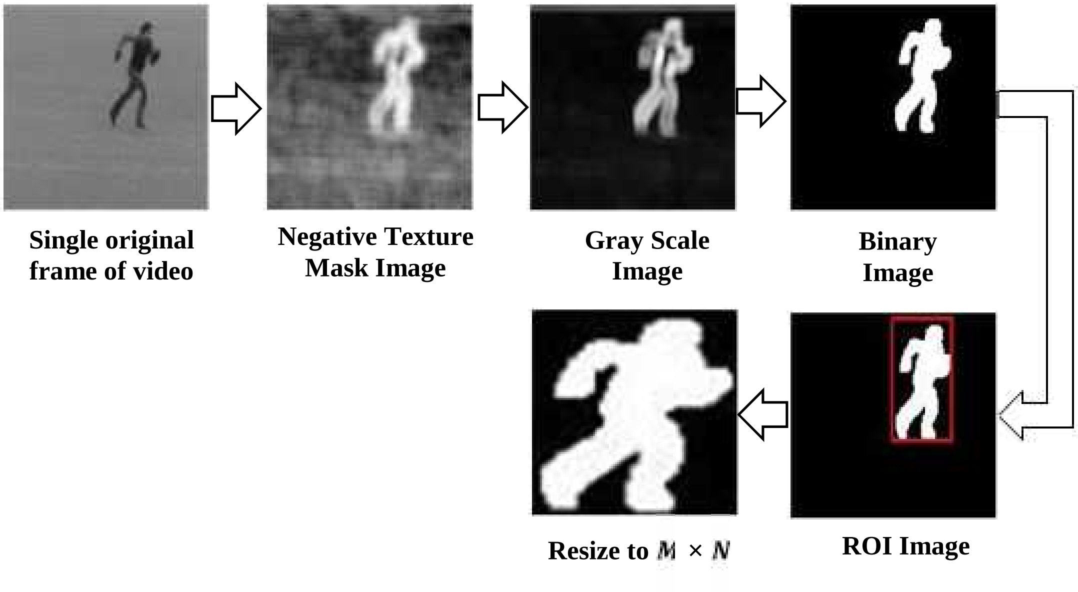


Figure 3.2: The flow diagram of silhouette extraction.

The segmented image may contain different white contours, but all of them are not human silhouettes. By comparing the size of these contours one can find the contour with the largest area which is fortunately, the human silhouette. This part of the image is selected and classified as human silhouette. As in Figure 3.2, two parts are shown that have the same texture, but human part has the larger area and therefore it has been selected as a silhouette.

3.2.2 Feature Extraction and Representation

Recently, it has been observed that the concept of visual recognition in static images [93] has been successfully extended to video sequences representation. The various methods [55] [94] [95], used for representing human activities are: the feature detectors, feature descriptors, bag-of-features representations, and local features based on voting for localization. The feature extraction is the prime step for analysis of video sequences and extracted features must be robust, and invariant against the variation of recording conditions, body pose, etc. All the analysis is done over the feature set collected and based on that, one can find the desired results by applying different techniques. Subsequent sections describe about the key pose selection, and feature extraction and representation.

3.2.2.1 Selection of Key Poses

In general some of the video frames do not contain any information content about the object. Consider a person who is 'walking' and the camera is still, he will pass in front of the camera for a short time, and most of the time frames do not have any content with 'human blob.' To select the frames, which have maximum information content, the key frames are extracted and are used for the purpose of feature extraction. Figure 3.3 depicts the mechanism that is involved in the extraction of key frames out of a large number of frames present in a video.

The k −key frames, which have significant energy value as compared to the highest energy value of the frame, are chosen for further processing, and energy of the frame is calculated using Equation 3.2. These k −key frames are kept in a timed sequence with reference to the highest energy frame and by this arrangement of key frames the spatial change of the shape with respect to time is maintained. Computation of these key frames is extremely robust in discriminating different human activity due to their ability to discriminate both spatial and temporal information. Additionally, selection of key poses improves the redundancy and also eliminate the geometrical loss of information. Hence, it becomes extremely important to select the key poses of the human activity for its effective representation.

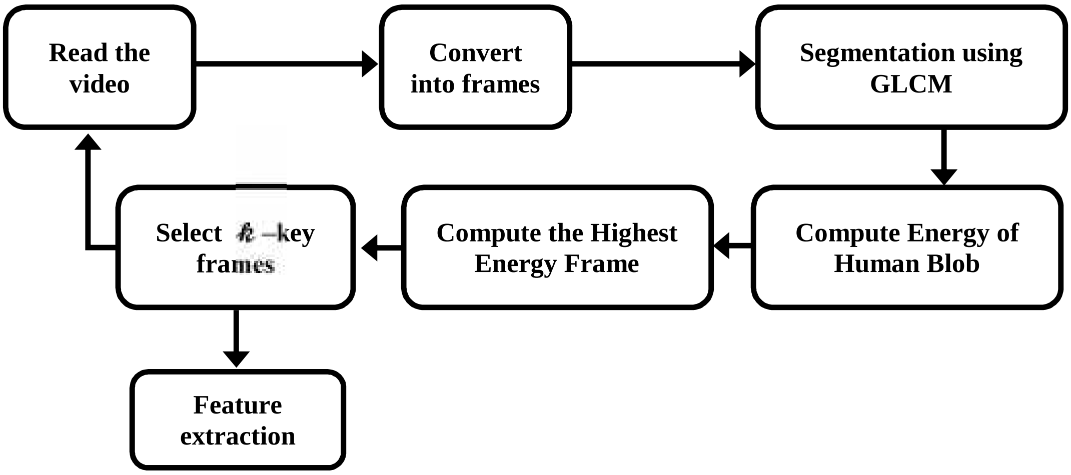


Figure 3.3: Flow diagram of selecting key poses frames.

The timed sequence silhouette frames are divided into a number of cells, and each cell contains different number of white pixels. To maintain uniformity, the size of the frames has to be fixed. The difference in the size of the obtained silhouette may give different information and may lead to misclassification or error. The silhouettes of person extracted from earlier steps are not of the uniform size and therefore, it is necessary to resize the images.

3.2.2.2 Cells Formation

The resized image of $M \times N$ contains total pixels equivalent to ℕ. It is divided into a grid of $\mathcal{U} \times \mathcal{V}$ images as depicted in Figure 3.4. Since the image is converted into binary form, therefore the white pixel can be calculated in the cell and the number of white pixel is used as a feature for this particular cell or grid. Similarly, the number white pixels in each cell is computed to form the feature vector.

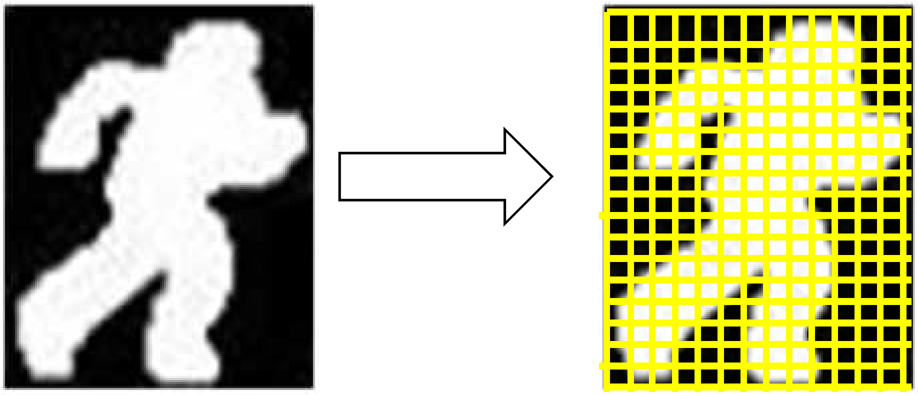


Figure 3.4: Formation of cells using grid.

3.2.2.3 Procedure of Feature Extraction and Representation

Consider a segmented video of an activity that contains a finite number ($\mathfrak{N}$) of frames, represented as $I_t(x,y)$, where t represents the frame number i.e. $t \in \{1,2,3,\dots,\mathfrak{N}\}$ and x, y are the dimensions of the frames. To maintain uniformity, the next step is to resize each frame to a size of $M \times N$.

For effective and efficient representation of activity, only key poses of the frames are chosen and these are the frames that have higher energy in the video sequence compared to other frames. The energy of a frame is calculated as:

$$U_t = \sum_{i=1}^{M} \sum_{j=1}^{N} \|I_t(i,j)\|^2 \qquad (3.2)$$

For the selection of key poses of the silhouette frames, a sequential search operation is applied up to a certain number of frames to find highest energy value of silhouette frame among all frames $(U_1, U_2, U_3, U_4, \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots . . U_{\mathfrak{N}})$. The highest energy value frame is considered as the reference frame and the frames which have significant energy value as compared to the highest energy value frame are selected as $\Bbbk$-key frames of the silhouette. Now, each key frame is divided into cell images $C_i(x, y)$, where each cell is of size $u \times v$, and the total number of cells therefore are calculated as:

$$\frac{M}{u} \times \frac{N}{v} = N_c \tag{3.3}$$

where N_c is total number of cells in the key frames and denoted as $(C_1, C_2, C_3, C_4, \dots . . . , C_{N_c})$.Since silhouette images are binary images, they may contain only white or black pixels and the number of white pixels are counted as:

$$w_i = count\{C_i(x, y)\},\ where\ i = 0,1,2, \dots \dots \dots \dots . N_c. \tag{3.4}$$

Where w_i represents the number of white pixels in the *i*th cell, the counted number of pixels in a cell of one frame are arranged in such a manner that retains the time sequence of the action and can be represented as:

$$f_i = \{w_1, w_2, w_3, w_4, \dots \dots \dots \dots \dots w_{N_c}\}, where\ i = 1,2,3, \dots \dots \dots \Bbbk. \tag{3.5}$$

where f_i contains the white pixel count of the *i*th key frame. Thus, the feature vector of an activity video sequence is expressed as:

$$F_i = [f_1, f_2, f_3, f_4, \dots \dots . \dots \dots \dots . f_{\Bbbk}], where\ i = 1,2,3,4, \dots \dots \dots \dots . . V_T. \tag{3.6}$$

where V_T is the total number of videos in the data set. Substituting the Equation 3.5 in Equation 3.6, then feature vector for a single video of activity is:

$$F_i = \left[\underbrace{w_1, w_2, w_3, \dots \dots w_{N_c}}_{First\ Frame}, w_1, w_2, w_3, \dots \dots \dots \underbrace{w_1, w_2, w_3, \dots \dots w_{N_c}}_{Last\ Frame\ i.e.\ k^{th}} \right] \tag{3.7}$$

Similarly, the feature set of a dataset which contains V_T number of videos of all classes can be represented as:

$$F_v = \begin{pmatrix} F_1 = & \underbrace{w_1, w_2, w_3, \dots \dots w_{N_c}}_{First\ Frame}, w_1, w_2, w_3, \dots \dots \dots, \underbrace{w_1, w_2, w_3, \dots \dots w_{N_c}}_{Last\ Frame\ i.e.\ k^{th}} \\ F_2 = & \underbrace{w_1, w_2, w_3, \dots \dots w_{N_c}}_{First\ Frame}, w_1, w_2, w_3, \dots \dots \dots, \underbrace{w_1, w_2, w_3, \dots \dots w_{N_c}}_{Last\ Frame\ i.e.\ k^{th}} \\ & . \\ & . \\ & . \\ & . \\ F_{V_T} = & \underbrace{w_1, w_2, w_3, \dots \dots w_{N_c}}_{First\ Frame}, w_1, w_2, w_3, \dots \dots \dots, \underbrace{w_1, w_2, w_3, \dots \dots w_{N_c}}_{Last\ Frame\ i.e.\ k^{th}} \end{pmatrix} \tag{3.8}$$

where v corresponds to the videos of all classes. One of the important aspects of the feature set is its dimension. If the dimension is high, then it must be reduced to a lower dimension for efficient and speedy classification. The dimension of the feature set F_v can be determined by determining the dimension of feature vector representing a single video in Equation 3.7. It is determined to be $N_f = N_c \times k$ and after concatenation the dimension is written as $1 \times N_f$. Hence, the dimension of the final feature set F_v is determined as $V_T \times N_f$. Finally, in order to recognize actions, these feature vectors and their labels are passed to the classifiers.

3.2.3 Classification Models

Usually, the computed data of feature set have the correlated and uncorrelated data together and the correlated feature set of different classes makes the classification complex and slow. Hence, for efficient and fast classification, the feature set must have uncorrelated data since the uncorrelated feature set has a lower dimension. So, for easy handling of data and improved classification, the dimension of feature set must be reduced. The PCA [81] is a popular method for the reduction of dimension of the feature set by maximizing the variance of the feature set and mapping the feature set into a lower dimensional space.

The significant lower dimension data is classified by training and testing of the feature set. There are two types of classifiers used for the human activity recognition, which are i) Linear and ii) Nonlinear. We have used one linear and two nonlinear classifiers. The linear classifier is the LDA, and nonlinear classifiers are: K-NN, and SVM. In addition to these one more nonlinear classification model is proposed using the combination of SVM and K-NN.

Originally, the concept of Linear Discriminant Analysis (LDA) was proposed by [83], also known as Fisher’s Discriminant Analysis. It provides an improved classification of various classes of data by maximizing the margin between dissimilar classes and minimizing the margin within the same classes.

The K-NN classifier [84] selects the K-closest samples of training feature set to the new instance, and the nearest class having highest voting is mapped to test instance. One of the biggest advantages of this classifier is its non-parametric nature. It does not require any assumptions and easily classifies the data even with a higher dimension space.

Support vector machine (SVM) is a machine learning classifier based on the structural risk minimization principle [96]. It is one of the widely used classifiers for classification of human activity [87]. The term support vectors are features of the samples, which are close to the hyperplane of an SVM. Therefore, the determination of the location of most important data is near the hyperplane and it is formed by the set of lines drawn between the class samples.

3.2.3.1 Hybrid Classifier

To enhance the classification accuracy, a hybrid classification model is developed, which comprises SVM and K-NN classifier. As it is shown in Figure 3.5, the support vectors (encircled) are samples of feature set, which are close to the decision boundary and play a decisive role in the classification of a test sample. If a particular test sample is close to the hyperplane, then an accurate classification is not always possible, and SVM may classify these sample inaccurately. Therefore, in this classification model, the wrongly classified samples are further classified by clustering these samples.

It is also vital to mention here that the wrongly classified test samples through SVM, which are close to the hyperplane, become the representative points for K-NN. These wrongly classified samples are further classified using the 1-NN classifier, and if the 1-NN correctly classifies any wrongly classified sample of SVM then the original classification result obtained by SVM is updated, otherwise it remains unaltered. Hence, by use of multiple classifications, the overall classification accuracy can be improved.

The performance of the hybrid classifier is evaluated using model of HAR system whose block diagram is as shown in Figure 3.6, where the dotted box is the hybrid classifier model. The proposed model is the partial cascading of two classifiers (SVM, KNN), where the human activity feature sets are extracted from human silhouettes and subsequently, these sets are dimensionally reduced by PCA. Further, the reduced feature sets are used for training and testing of the SVM classifier. The test scheme used for the classification of these feature sets is leave-one-out cross-validation (LOOCV).

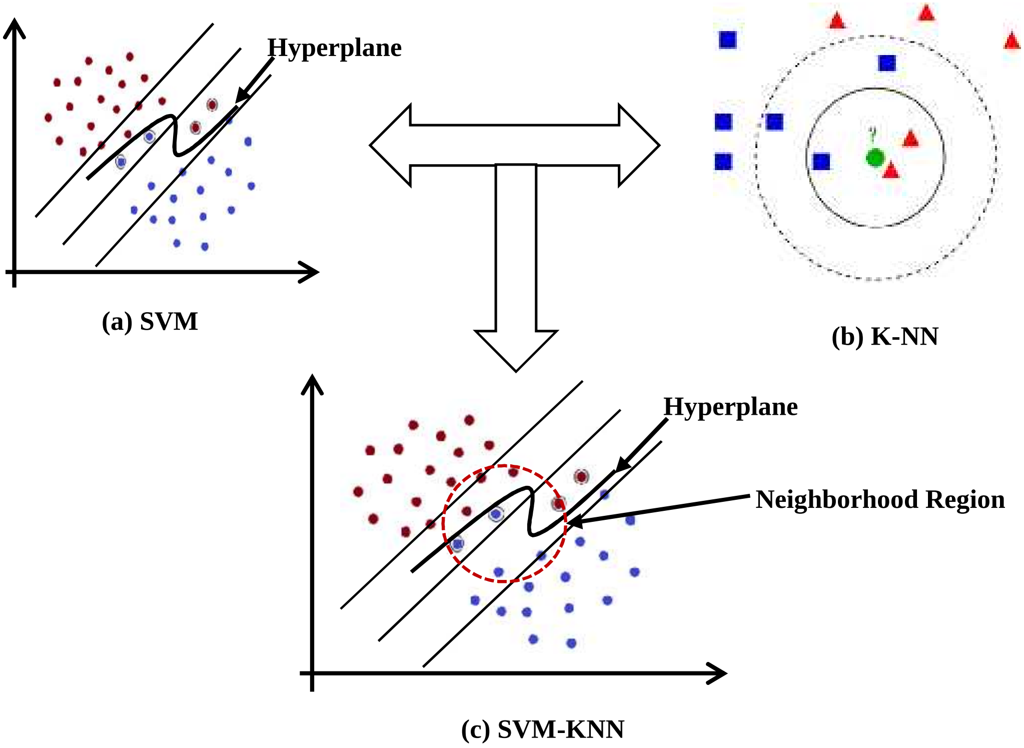


Figure 3.5: The formation of hybrid classifier (SVM-KNN) in feature space.

In the LOO scheme, the samples of one class are input together, and the result is predicted by the classifier based on the training. If there are N-number of classes, then the testing is performed N times, and a confusion matrix of N×N order is formed and passed to the fuser unit. The confusion matrix consists of the predicted class samples against the input class samples in the form of true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

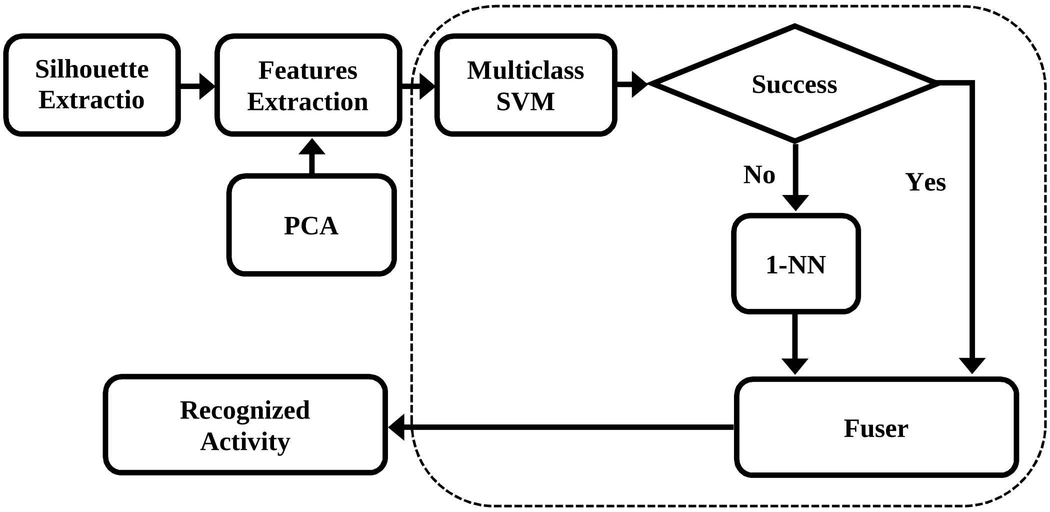


Figure 3.6: Hybrid classification model for HAR.

The correctly predicted sample are stored as TP and wrongly predicted sample are stored as FN. The FN samples of each class are further classified using the 1-NN classifier, and if any sample of FN is classified as a TP sample category, then the original confusion matrix having TP sample values of each class is updated by the fuser unit, else it remains unchanged. Hence, in this way the overall accuracy is improved. The detailed performance analysis of the hybrid classifier is presented in the subsequent section.

3.3 Experimental Work

In order to assess the effectiveness of the proposed approach, we have conducted experiments on three public benchmarks, the Weizmann dataset [74], the KTH dataset [97], and the Ballet dataset [98]. The assessments include a range of variations resulting from different resolutions, lighting conditions, occlusions, background disorder, and uneven motions. In this experiment, for the representation of all videos of the data sets,

25 key frames of size 50×30 are used to represent an activity sequence and each frame has 60 number of cells, which is obtained by considering the size of each cell 5×5. Hence, the dimension of a silhouette is $25 \times 60 = 1500$, and after concatenation the silhouette feature vector is represented as 1×1500. To evaluate the outcome of action classifications, a leave-one-out (LOO) cross validation scheme is opted for all the data sets. For each data set, we have used three machine learned classifier (LDA, K-NN, SVM). In addition to these, to enhance the recognition accuracy a hybrid “SVM-NN” classification model is proposed. The average recognition accuracy (ARA) is computed using Equation 3.9.

$$\mathrm{ARA} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \times 100 \text{ (In percentage)} \qquad (3.9)$$

where TP, TN, FP and FN are the number of true positive, true negative, false positive, and false negative, respectively. The highest obtained recognition accuracy of these classifiers are compared with the similar state-of-the-art methods.

Weizmann data set: This dataset was introduced by Gorelick et al. [74], which consists of 90 videos with a frame rate of 25fps and each frame has a size of 144× 180. The sample frames of the data set are shown in Figure 3.7. In the video sequence, 9 people are performing 10 different actions, categorized as walk, run, jump-jack, bend, jumping forward on one leg, jumping on two legs in the forward direction, jumping in place, sideways jump, one hand wave, two hand wave. This is one of the established benchmarks of the data sets for the human action recognition and most of the earlier methods’ [58] [72] [74] [77] [85] [99] recognition accuracy is computed on this data set.

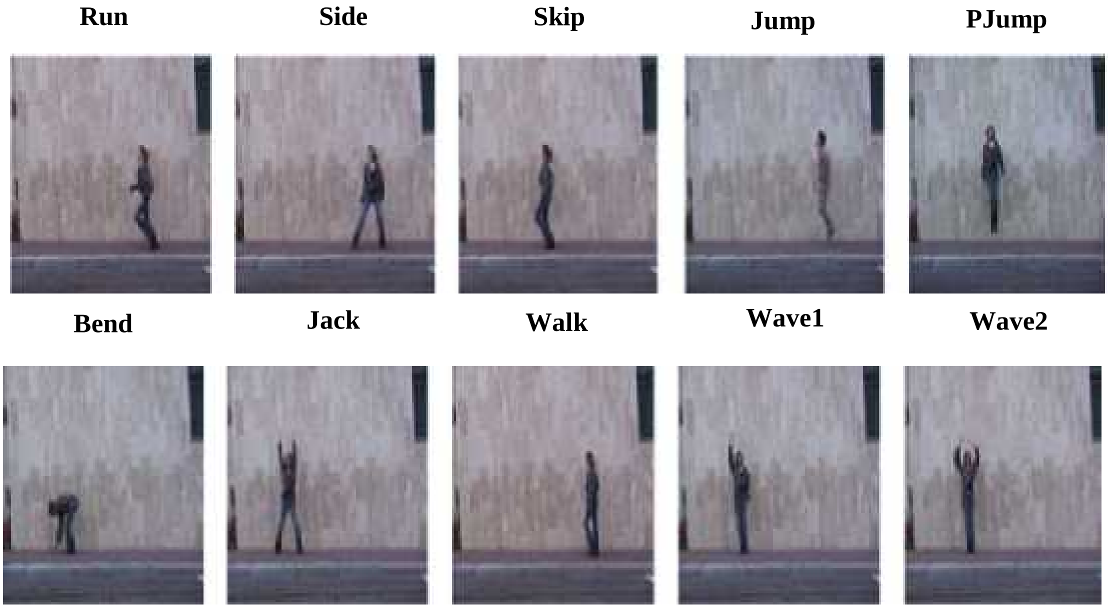


Figure 3.7: Sample frames of Weizmann human action dataset

KTH data set: This dataset was introduced by Schuldt et al. [97] and is a more challenging dataset as compared to the Weizmann dataset. The sample images of the data set are depicted in Figure 3.8. The dataset consists of six basic activities, namely; "Hand-Clapping," "Hand-Waving," "Jogging," "Jumping," "Running," and "Walking". Each activity has 100 videos for four different scenarios (S1, S2, S3 and S4) in different light conditions, indoor and outdoor conditions. All these video sequences are recorded in a uniform background with a static camera of frame rate 25fps and further down-sampled to the spatial resolution of 160x120 pixels. The recording conditions of the videos in the KTH data set are not stationary and there is a significant amount of camera movement and lighting effects in some cases. Therefore, the silhouette extraction is not straight forward and simple background subtraction method may not suitable. Hence, for the silhouette extraction we have incorporated the texture based segmentation method.

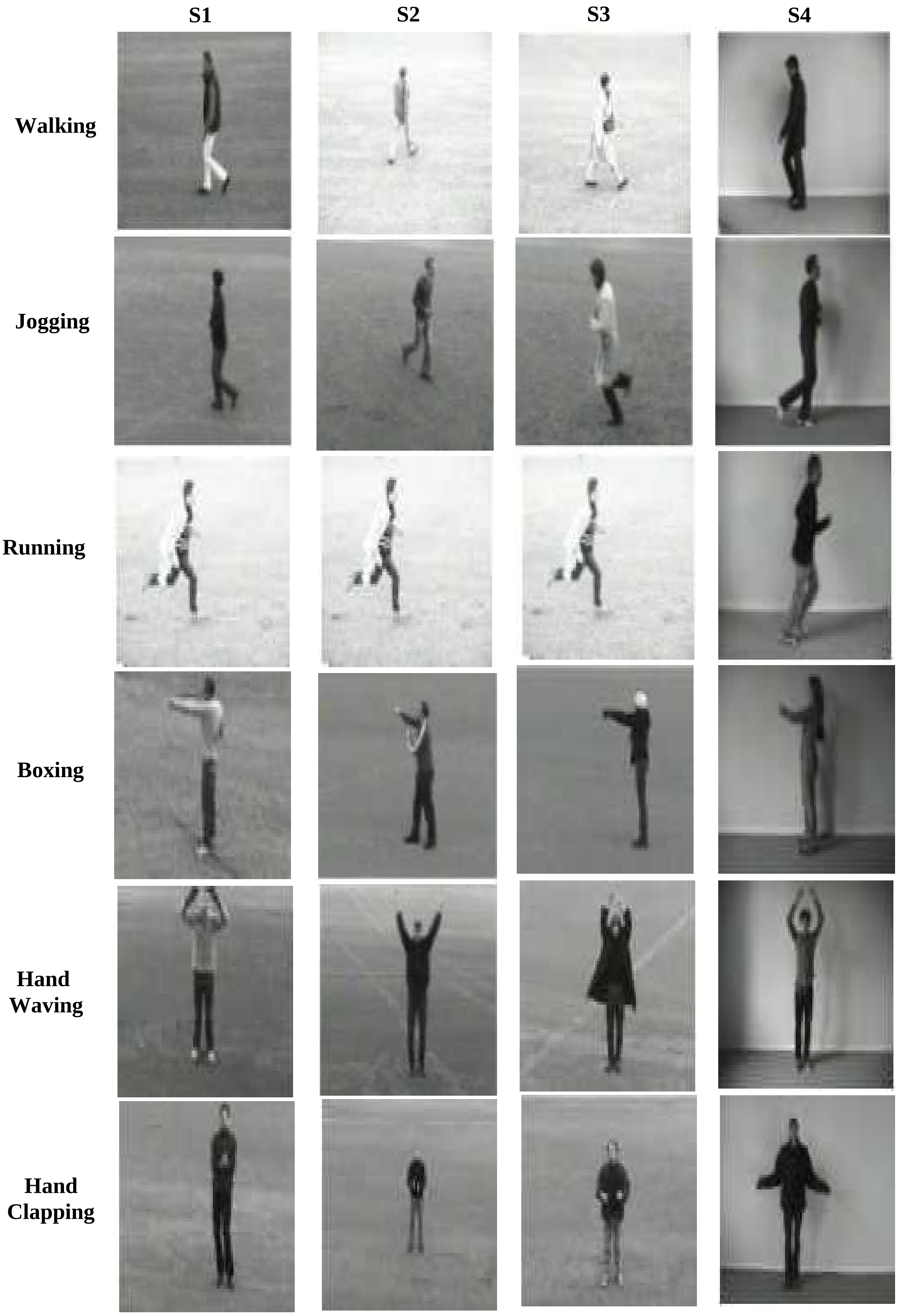


Figure 3.8: Sample frames of KTH dataset.

Ballet data set: The Ballet Movement action data set [98] is one of the complex human action data sets. The sample frames of the data set are depicted in Figure 3.9. This data set consist of eight ballet movements performed by three actors and these movements are named as “ Hopping (HP),” “Jump (JP),” “Left-to-Right Hand Opening (LRHO),” “Leg Swinging (LS),” “Right-to-Left Hand Opening (RLHO),” “Standing with Hand Opening (SHO),” “Stand Still (SS)” and “Turning (TR),”. The data set is highly challenging due to the considerable amount of intra-class dissimilarities in terms of spatial and temporal scale, speed, and clothing.

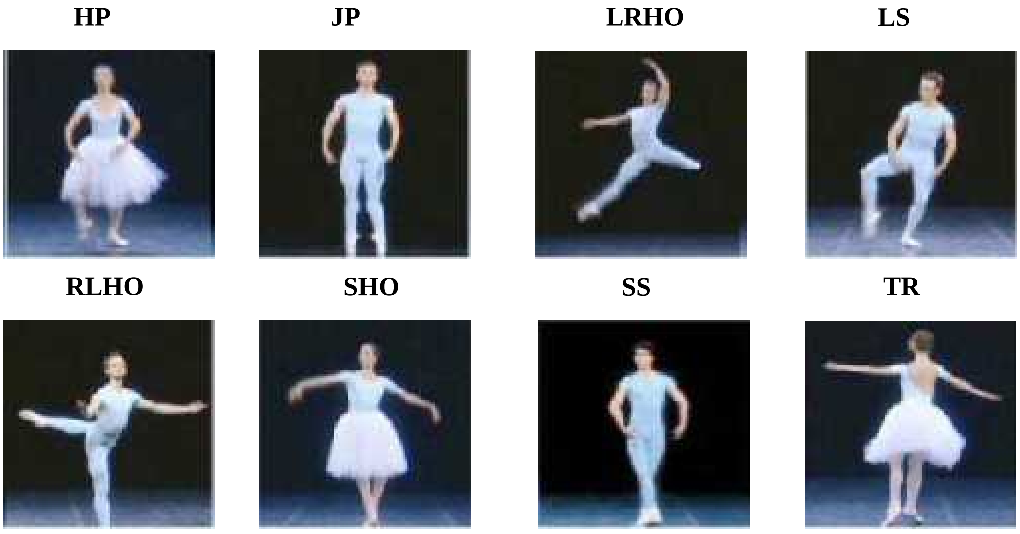


Figure 3.9 Images of the Ballet data set depicting eight movements of actions.

3.3.1 Classification Results

The classification results of the proposed approach are depicted in Table 3.1 on three different data sets using four different classification models, including the proposed one i.e. “SVM-NN”. The aim of this depiction is to show the effectiveness of the proposed descriptor as well as the performance of the proposed classification model in comparison with the existing models.

The classification strategy opted for the Weizmann data set is as per Gorelick et al. [74]. From Table 3.1, the recognition accuracy obtained by LDA is less as compared to the other classifiers. This is due to the similarity between the activities of the data set like running, jumping and walking and therefore activities are very difficult to separate using a linear model of classification. The highest ARA achieved in this experiment is through hybrid “SVM-NN” classification model which is 100%. The hybrid classifier is the combination of two nonlinear classifiers and these classifiers are more adept to inter class similarities and intra class dissimilarities, hence giving the improved result.

For the KTH data set, the variation in recording condition is more as compared to the Weizmann data set due to which the extraction of silhouette in this data set is a difficult task. Simple frame differencing methods for extraction of silhouette may give good results in Weizmann data set, but in the case of KTH, it is very difficult to extract silhouette using these methods. Due to the variation in recording conditions the texture of object is the least affected. Hence, texture based foreground extraction is utilized. The highest ARA achieved on the KTH dataset in this experiment is 96.4 % as shown in Table 3.1. From Table 3.1, it can also be seen that the performance of various classifiers on the KTH data set is increasing. For the Ballet dataset, the highest ARA achieved is 94% through hybrid classifier as shown in Table 3.1. The performance of this approach is least on this dataset as compared to the Weizmann and KTH datasets, because of the complex motion patterns, which differentiates the execution of the motion from actor to actor. The misclassification error is instigated by the "hopping" as it is confused with a much related activity "jump".

Table 3.1: Classification results on the datasets in terms of ARA (%)

Datasets\ Classifiers	**Weizmann**	**KTH**	**Ballet**	**mRA (%)**
LDA	94.4	90	75	86.4
KNN	96.6	91.7	90.2	92.8
SVM	97.7	92.4	92.75	94.28
Hybrid	**100**	**96.4**	**94**	**96.8**

3.3.2 Comparison of Recognition Accuracy

Comparison of recognition accuracy is carried out in two stages: first, the performance of proposed "SVM-NN" classifier is compared with others (LDA, KNN, and SVM) and the second stage, the highest average recognition accuracy obtained on the datasets, is compared with the methods of others.

The classifier performance is compared through the mean classification error (MCE), which is computed using Mean Recognition Accuracy (mRA). The MCE of all the classifiers is depicted in the Figure 3.10, and lowest classification error is found in the case of SVM-NN classifier. LDA is a linear classifier, the speed of processing of which is faster than a kernel based classifier, but the speed comes at the cost of efficiency. The MCE is 13.6 %, which is higher than others and hence it can be concluded that it may not be a good classifier for the dataset which has more interclass similarity. The MCE of a nearest neighbour classifier is almost 7.2 %, which is slightly lower than that of LDA and the reason for the lower MCE is the nonlinear nature of the classifier.

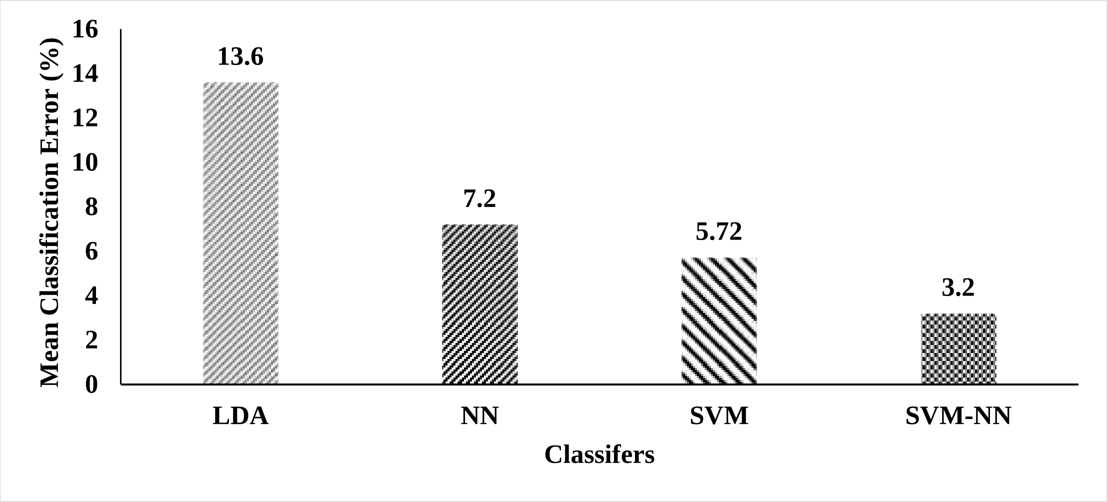


Figure 3.10: The comparative performance of the classifiers.

The SVM is a kernel based trick, and gives a unique solution because the optimality problem is convex. The MCE is reported as 5.7 *%*, which is lower than the LDA and KNN both. The individual performance of SVM and KNN approach for classification of human activity is already proven [58] [72] [74] [77] [85] [86] [99] [100] [101] [102]. Therefore, a hybrid “SVM-NN” classifier is proposed and the performance is measured in terms of MCE (3.2%), which is lowest among all the classifiers used.

The effectiveness of the proposed description methodology of human activity is analysed by comparison of the highest ARA achieved on each dataset with respect to the other methodologies as presented in Table 3.2, 3.3, 3.4. The highest ARA is achieved by the hybrid “SVM-NN” classification model on all the datasets used. The comparison is almost fair due to the similar conditions used in this experiment and others such as: input silhouette, classifier, test scheme and classifiers used. The classifier used in others technique are KNN and SVM.

Table 3.2 : Comparison of ARA on Weizmann dataset.

Method	Input	Classifiers	Test Scheme	ARA (%)
Gorelick et al. [74]	Silhouettes	KNN	LOO	97.5
Chaaraoui et al. [72]	Silhouettes	KNN	LOSO	92.8
Wu, & Shao [77]	Silhouettes	SVM	LOSO	97.78
Goudelis et al. [85]	Silhouettes	SVM	LOPO	95.42
Melfi et al. [58]	Silhouettes	SVM	LOO	99.02
Touati & Mignotte [99]	Silhouettes	KNN	LOO	92.3
Proposed Method	**Silhouettes**	**SVM-NN**	**LOO**	**100**

Table 3.3 : Comparison of ARA on KTH dataset.

Method	Input	Classifiers	Test scheme	ARA (%)
Sadek et al. [86]	Silhouettes	SVM	-	93.30
Saghafi & Rajan [100]	Silhouettes	KNN	LOO	92.6
Goudelis et al. [85]	Silhouettes	SVM	LOPO	93.14
Melfi et al. [58]	Silhouettes	SVM	LOO	95.25
Rahman et al. [101]	Silhouettes	KNN	LOO	94.49
Conde & Olivieri [102]	Images	KNN	-	91.3
Proposed Method	**Silhouettes**	**SVM-NN**	**LOO**	**96.70**

Table 3.4 : Comparison of ARA on Ballet movement dataset.

Method	Fathi & Mori, [98]	Wang & Mori [103]	Guha, & Ward [104]	Ming et al. [105]	Iosifidis et al. [106]	**Proposed Method**
ARA (%)	51	91.3	91.1	90.8	91.1	94

Tables 3.2 & 3.3, show the comparison of result achieved through hybrid classification model with the similar state-of-the-art methods on Weizmann and KTH datasets, respectively. The test methodologies used in these methods are Leave-One-Out (LOO),

Leave-One-Person-Out (LOPO), and Leave-One-Sequence-Out (LOSO), which are fairly similar to each other. Hence, the comparison on these two datasets is fair, because the experimental setup used in these techniques is similar to that in the proposed one. As it is seen in Table 3.2, the ARA of 100% is achieved on Weizmann dataset, which is higher than the other methods that use the SVM and KNN classification models. The reason for this high accuracy is the quality of silhouette extraction and effective representation, and capability of classifier to deal with intraclass variation among the activities. Similarly, as it is seen in Table 3.3, the ARA is 96.7%, which is again higher than the other methods that use the similar classification model.

Table 3.4, gives the comparison of result achieved through this approach with that of five earlier works which use the Ballet dataset. We have used the similar experimental setup as used in [98] [103] [105] [106] thus this comparison is fair. The work of Guha & Ward [104] uses different experimental setup as compared to one used in this experiment. Hence, this comparison may not be a right one, but given the complexity of the database, a higher ARA is obtained successfully by this technique, which is heartening.

3.4 Significant Findings

This chapter addresses the problem of low recognition rate and loss of geometrical information in the bags of words model of human activity recognition using key poses of human silhouette. The experimental result demonstrate the few interesting observations, which are as follows:

- The proposed approach of activity representation and classification model significantly outperforms many of the existing silhouette based activity recognition methods as can be seen from Table 3.2, 3.3, and 3.4.

- The classification model performs better than the LDA, KNN, and SVM as can be seen by Figure 3.9, in which the least error is achieved through SVM-NN.

- It can be seen that improvements achieved for the KTH and Ballet dataset are more significant because these two datasets have challenging environmental conditions and significant intra-class variations in terms of speed, spatiotemporal scaling, zooming in, zooming out, clothing, etc. and these are directly related to the input data. The silhouette extraction on Weizmann dataset is comparatively easy and accurate as compared to the KTH and Ballet dataset due to less variations in recording conditions.

- As the number of key poses increases, the complexity also increases, and it does not give a significant increase in recognition accuracy. On the other hand increasing the number of cells, gives marginal increase in the effectiveness but results in a higher dimension.

This chapter is based on the following work:

D.K. Vishwakarma, Rajiv Kapoor, “Hybrid classifier based human activity recognition using the silhouette and cells”, ***Expert Systems with Applications*** (**Publisher: Elsevier**), 2015, Vol. 42, No. 20, pp. 6957–6965. [29].

CHAPTER 4

SPATIAL DISTRIBUTION AND ROTATION ANALYSIS

The objective of this chapter is to describe the human activity recognition approaches based on the computation of spatial distribution of gradients of average energy silhouette images and rotational analysis of human silhouette frames. The key ingredients in this chapter include the exhaustive study of computation spatial distribution of intensities in an average energy silhouette images (AESIs), sum of directional pixels computation, computation of rotation using □-transform, generation of feature map model, experimental works, results discussion and comparative analysis of results with the similar state-of-the-arts.

4.1 Introduction

The spatial distribution of intensities in a gray scale image is the key constituent which represents the shape of an object. The distribution of intensities in gray scale AESIs varies according to the motion of human body parts and due to motion of the human body, the orientation human silhouette also changes over the time. The degree of rotation is different for different human activities. Based upon these features of the human activities, the recognition approaches can be developed by the computation of shape and rotation information. The shape of human activities are computed by measuring the spatial distribution of gradients (SDGs) and sum of directional pixels (SDPs) of AESIs. The rotation of human binary silhouette frames are computed using □–transform. For the computation of SDGs of AESIs, two approaches are developed.

The first approaches is based on pyramid of histogram oriented gradients and the second approach is based on the analysis from the exhaustive study of pyramid of histogram oriented gradients. Based on these two approaches of spatial distribution computation of intensities, the HAR models are developed by fusion of SDGs, SDPs and □–Transform. The key steps of these approaches are the extraction of human silhouettes, formation of average energy images, feature extractions using SDGs, SDPs and □–transform, formation feature map model, and classifications. The extraction of human silhouette from the video sequence is done using texture segmentation approach as described in Chapter 3. The formation of AESI is done through the summation of binary human silhouette frames, divided by the number of silhouette frames. The computation of SDGs are performed by measuring the magnitude and gradients of each pixel of a gray scale AESI as a whole and different decomposition levels of AESI. The SDPs are computed in the x and y direction of the AESI. The rotation of human silhouette images is computed using □–transform. These computed features are fused together to form a new descriptor. The fusion of these features possesses the advantages of both local and global features of the silhouette and thus provides an effective feature for representation model for HAR system.

Recently, several researchers [30] [31] [32] [33] [79] [80] have advocated that the multiple features based fusion technique can provide better performance than the individual features for the recognition of human actions and activities. Therefore, inspired from these approaches, a two novel integrated structure is developed based on the combination of appearance as well as the orientation of the human body silhouette. Another important fact behind use of multiple features is the human action dynamics, which states that in general a human activity is composed of translation and rotation of the human body. Due to the translation and rotation of human body, the shape of the silhouette changes with respect to time and in this approach the shape and rotation of human silhouette is computed and clubbed together at the recognition stage. The clubbing of features also give the local and global information. The main motivation is to formulate rich feature vector vocabulary having both the appearance and angular kinematics information that overcomes the limitation of earlier approaches. The pose dictionary formed, thus yields the human appearance representation and sequence of orientation provides the nature of the activity.

The performance of these approaches is tested under challenging publically available datasets i.e. Weizmann, KTH, and Ballet movements. The highest recognition accuracy achieved on each dataset is compared with similar state-of-the-art which demonstrate superior performance. The detailed description of developed approaches, experimental results and discussion of results are described in the subsequent sections.

4.2 Approach based on SDGs, SDPs and ℜ-Transform

In this approach, the multiple features of human silhouette are computed and these are SDGs, SDPs, and □- transform. The spatial distribution of gradients of AESIs are computed by calculating the magnitude and gradients of each pixels at various decomposition levels. The second feature is the SDPs of AESI, which is computed by determining the sum of intensities of the pixels in x-y directions. The third feature is the rotation of the human silhouette, which is computed through the □- transform. The first and second features give the shape of the human activity, while the third feature gives the rotation of the human activity due to motion. The prime requirement for the computation of these features is the accurate extraction of human silhouette frames from the video sequence, which is done using texture based segmentation approach. The overview of the proposed model is as depicted in Figure 4.1.

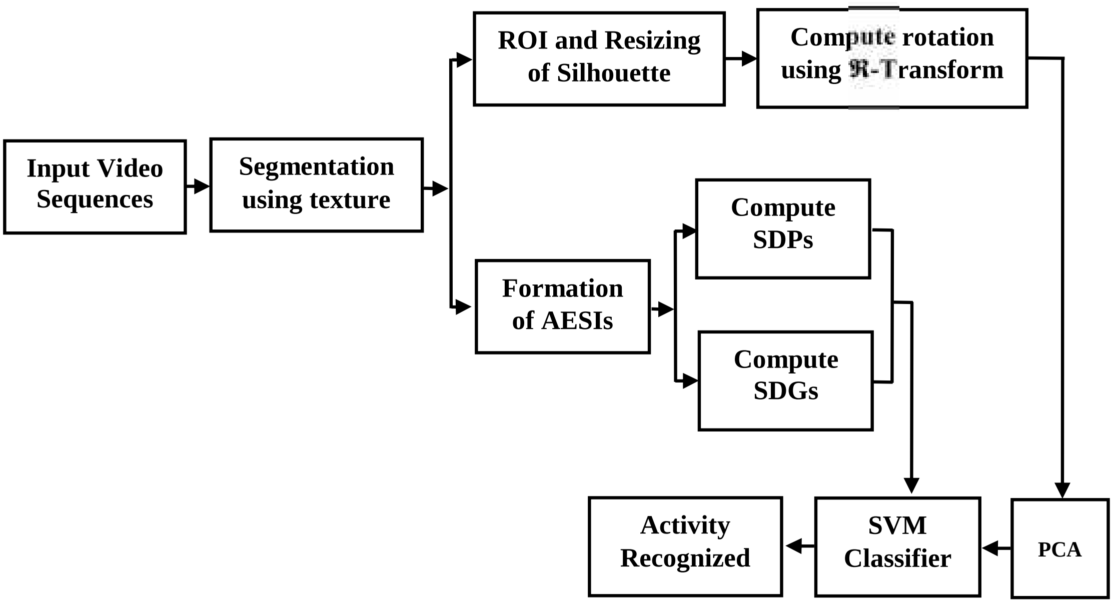


Figure 4.1: Work flow diagram of proposed methodology.

In Figure 4.1, the input of the video sequence is the video feed of human activity, which is inputted through a CCTV or a video camera or any other vision device capable of recording the event. The input video signal can be either in colour or gray scale spaces and depending upon the quality of video feed, some pre-processing operations (Enhancement, de-noising, colour space conversion etc.) may be required. The segmentation of binary human silhouette of activity in the video sequence is done using texture based segmentation procedure. The detailed description of segmentation is given in Chapter 3.

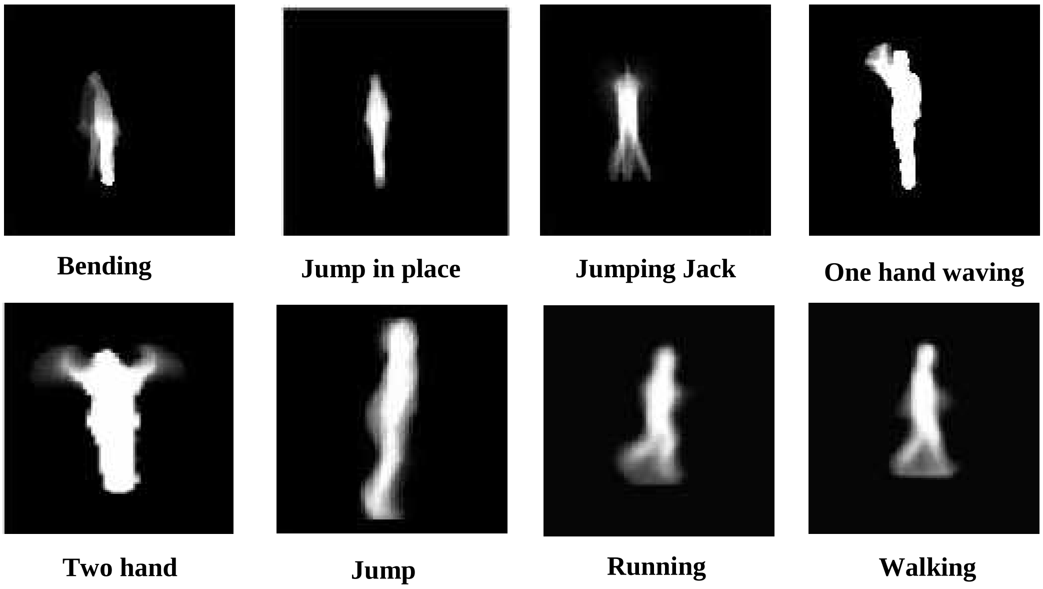
The SDGs and SDPs are used as local descriptors for the computation of structural information of AESI. On the other hand, the □-Transform gives the rotation based global information of the human silhouettes.

4.2.1 Average Energy Silhouette Image

The average energy silhouette image (AESI) is formed by the summation of binary human silhouette frames, divided by the total number of frames of human activity in cycle. The formation of AESIs is based on the accurate extraction of human silhouette, which is a prerequisite entity. As discussed in Chapter 3, the extraction of human silhouette can be done in many ways but in this approach, it is done using texture based entropy model, which is one of the most reliable and effective method of accurate extraction of silhouette under the noise and illumination changes [107]. A quite similar approach for the representation of human motion using Gait Energy Images (GEIs) was introduced by Han and Bhanu [108], in which they used to align the silhouette horizontally with respect to horizontal centroid, but in this approach, the alignment of the binary silhouettes is done with respect to a reference point and then the sequence is compressed into a single image based on incremental procedure, which takes into account the changes of subsequent silhouette frames. The AESI is mathematically defined as:

$$A_E(x,y) = \frac{1}{n}\sum_{\tau=1}^{\zeta}|\psi_\tau(x,y)|^2 \tag{4.1}$$

where, 'n' is the number of frames in a complete cycle, '$\psi_\tau(x,y)$' is defined as binary silhouette frame at time instant 'τ' and x, y are the pixel values in 2D image coordinate. The benefit of using these images is that it eliminates the time dependency restrictions [71] and effortless discrimination of similar representation of activities of different classes on the basis of intensity of the pixel values.



Bending Jump in place Jumping Jack One hand waving

Two hand Jump Running Walking

Figure 4.2: Representation of AESIs of different activities.

From the Figure 4.2, it can be observed and inferred that AESIs are just a reflection of 3D postures of human activity into 2D postures. The variation of intensities in these images signifies the motion of human body. The region, where more variation is taking place means that particular part of body is highly in motion as compared to the rest of the body part such as the two hand waving activity has more variation, in the upper body part where hand is moving. Similarly, the interpretation made for the slow variation or constant intensities means that part of human body is slowly moving or stationary such as the tummy of the human body is considered as the most stationary part as compared to the rest of the body parts while a person is performing an activity. Therefore, the region where tummy of human body is located, the image intensities are almost constant. The human activities or actions are performed by the movement of different body parts and their positions in the human body are also at different place. Hence, the AESIs shown in the Figure 4.2 are different from each other with respect to their intensity distribution. In some cases, where the AESIs have less and quite similar variations of intensity of different activities, indicate that only AESIs based representation of human activity is not enough for accurate representation of the human activities due to less temporal information. Therefore, in this work additionally, the motion based temporal features of the activities are computed using □-Transform and further integrated with the AESIs based representation of SDGs and SDPs. The process of computation SDGs, SDPs and □-Transform are as explained in the subsequent sections.

4.2.1.1 Computation of SDGs

The spatial distribution computation of edges of an object is used to categorize the class of an object based on the shape, presented by Bosch et al. [109]. In the similar manner, the SDGs features of AESI are computed by selecting the region of interest (ROI) of the AESI, which act as local descriptor. The selection of ROI provides more localized structural and noise free information and also ROI helps in creating these images invariant to scale and translation. The spatial distribution of an image $A_E(x,y)$ is computed by determining the magnitude and orientation of each pixel in the image, which is as shown in Figure 4.3. The magnitude and orientation of each pixel is determined using Equation 4.2:

$$\mathcal{M}(x,y) = \sqrt{A_{E_x}(x,y)^2 + A_{E_y}(x,y)^2} \quad \text{and} \quad \theta(x,y) = \arctan\left[\frac{A_{E_y}(x,y)}{A_{E_x}(x,y)}\right] \qquad (4.2)$$

where $\mathcal{M}(x,y)$ and $\theta(x,y)$ are the magnitude and gradients of the pixel respectively. $A_E(x,y)$ is the average energy silhouette image. Based upon the computation of magnitude and gradients of each pixels, the orientation bins are defined in the range of angle $(0-180^0)$ or $(0-360^0)$ [110], which is as shown in Figure 4.3. In the Figure 4.3, the 8-orienation bins are considered, where an image is given and each pixel of an image has different orientations in the range of $(0-360^0)$ but these pixel's magnitude in different orientations are quantized into 8-orientation bins for ease representation. The decomposition level of spatial distribution of gradient is defined as the uniform splitting of an image into different segments of an image in equal size as shown in Figure 4.3 and these segments are denoted as region (1), region (2), region (3), and region (4). Example: at level $(v=0)$, whole image is considered for the computation of spatial distribution of gradients, at level $(v=1)$, the input image is split into 4 sub-regions and spatial distribution of gradients is computed in each region, and at level $(v=2)$ image is split into 16 regions and spatial distribution is computed. Similarly, the process can be extended for v level of decomposition, where input image is divided into 4^v regions and SDGs features are computed in each regions. The final feature vector of SDGs computations is defined as: $\sum_{i=0}^{v-1} 4^i$, where $i = 0, 1, \dots . v-1$. The dimension of final SDGs feature vector is depends on the number decomposition level and number of orientation bins.

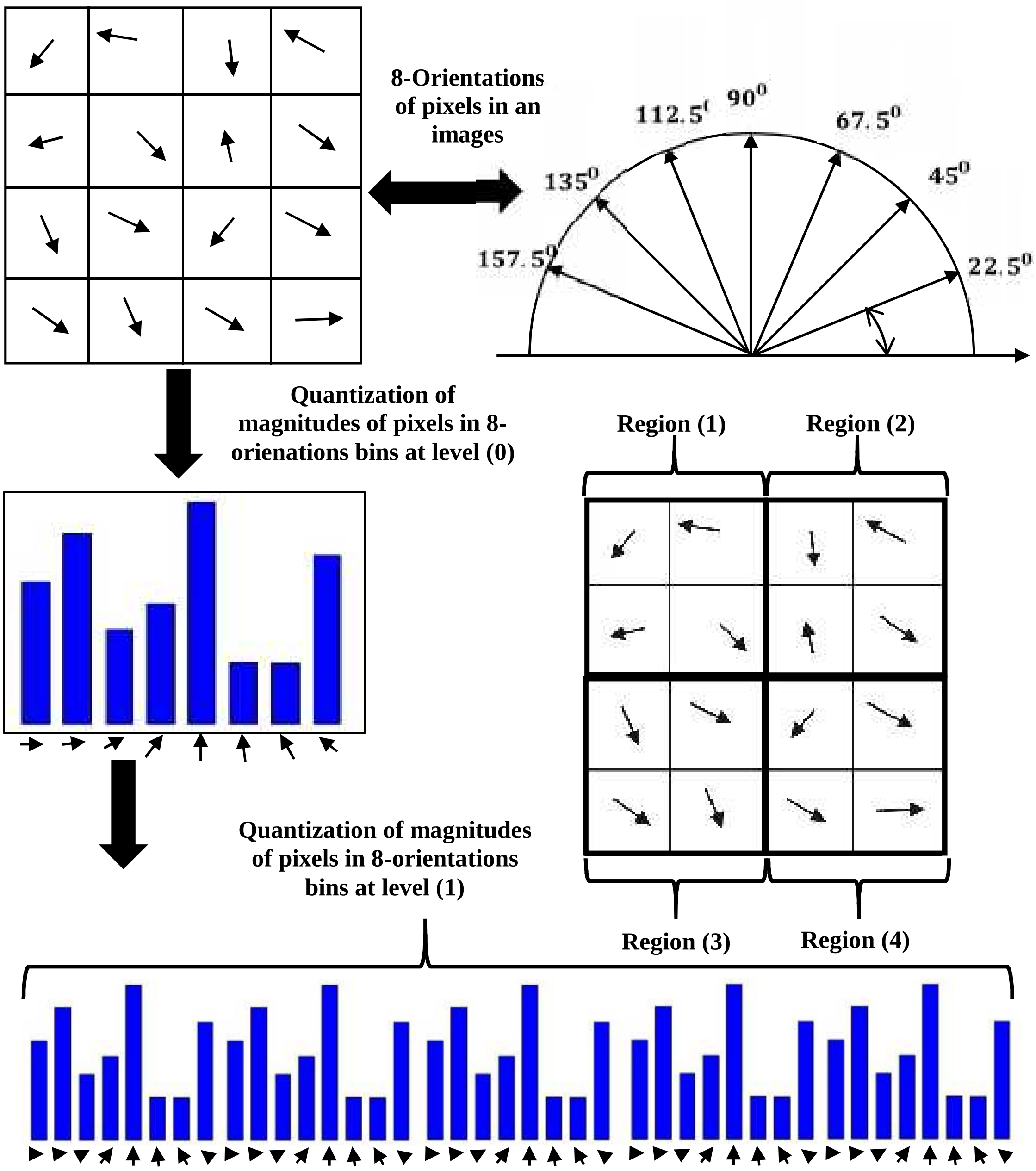


Figure 4.3: Depiction of SDGs formation at levels (0, 1) into 8-orientation bins.

The process of computing SDGs feature vector is as explained in Algorithm1. The input of the algorithm is the average energy silhouette image, the process of obtaining AESI is as explained in previous section.

```
Algorithm 1: Computation of SDGs

Input: Input an AESI $A^E(x,y)$.
Step1: Crop the ROI of AESI and denoted as $R_E(x,y)$. To maintain the uniformity,
        the dimension is resized into $M \times N$.
Step2: Compute the magnitude and gradient of each pixels of $R_E(x,y)$ image at
        different decomposition levels ($v$) using Equation 4.2.
Step3: Quantize the computed magnitude of intensity at various levels $(v)$ into $K$-
        orientation bins.
Step4: Adding all the regions together at decomposition level ($v$) and the final
        SDGs feature vector is generated.
Output: The histogram of feature vector is as shown in Figure 4.3.
```

The computation of magnitude and gradients in an average energy silhouette image is done using Equation 4.2. The process of quantization of the computed magnitude is done as shown in Figure 4.3, where the bar of histogram in the figure represents the bin that at a particular angle in the range of $(0 - 180^0)$ or $(0 - 360^0)$. These bins are placed at a uniformly interval of angle. The SDGs feature vector of different human activities are as shown in Figure 4.4., where it can clearly observed that the activities of different types leads to the different histogram. The dimension of histogram is 1×168, which means that the number orientation bins and level of decomposition are 8 and 2. The dimension of 1×168 is computed as $8 \times [4^0 + 4^1 + 4^2] = 168$.

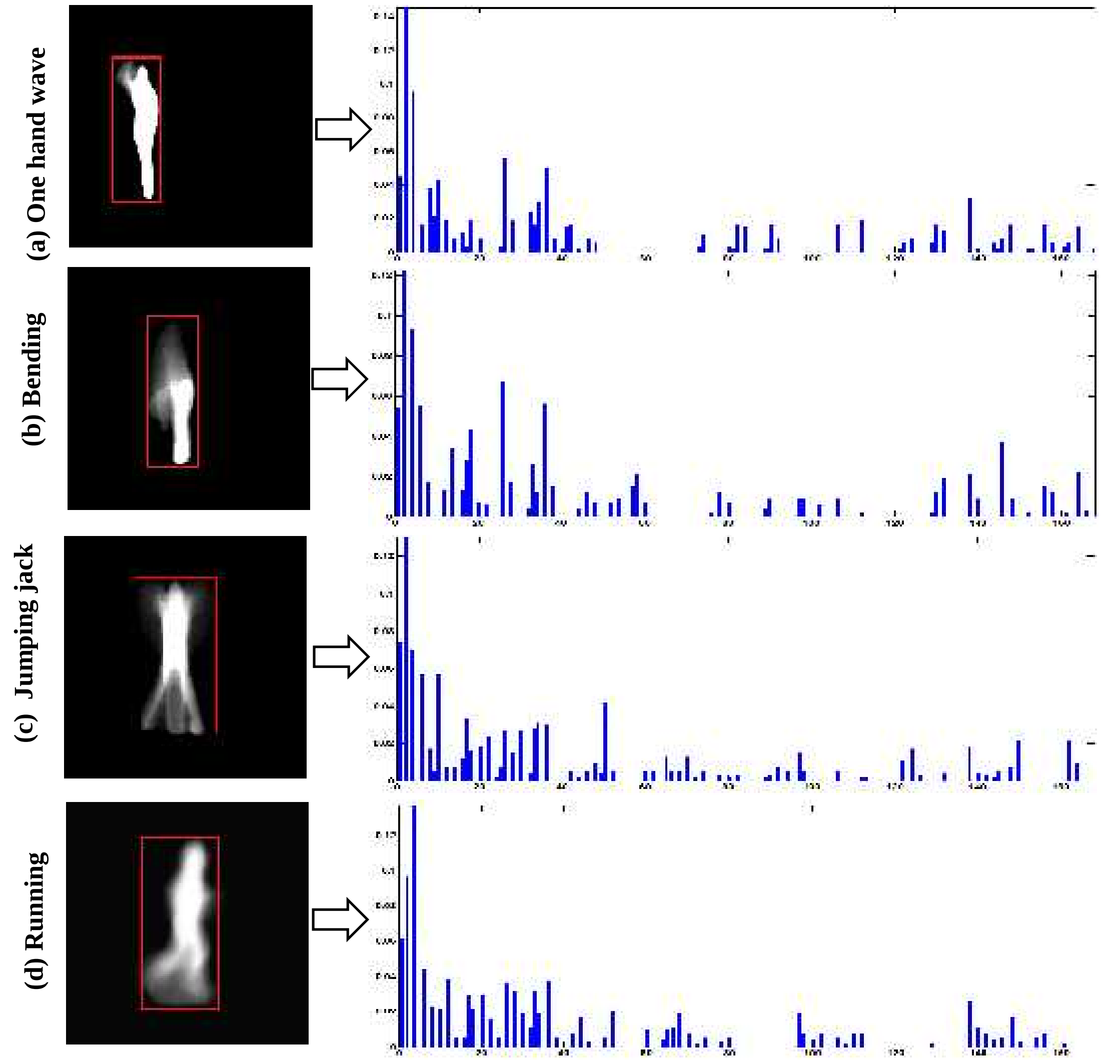


Figure 4.4: SDGs Representation at decomposition level-2 of different activities.

4.2.1.2 Computation of SDPs

SDPs of an image containing a particular geometrical shape can be represented by computing the pixels sum in horizontal and vertical directions of an image. The sum of directional pixel in horizontal and vertical directions are computed using Equation 4.3 and 4.4 and these equations are defined for an image of size $M \times N$ as follows:

$$H_X(k) = \sum_{l=0}^{N-1} \frac{A_E(k,l)}{\max(A_E(k))}, \quad 0 < k < M-1 \tag{4.3}$$

$$V_Y(l) = \sum_{k=0}^{M-1} \frac{A_E(k,l)}{\max(A_E(l))},\ 0 < l < N - 1 \tag{4.4}$$

where $H_X(k)$ and $V_Y(l)$ are the horizontal and vertical sum of pixels respectively. These pixels values are normalized by dividing by the maximum pixel value of the bin of each row for horizontal SDPs and each column for vertical SDPs computation. The normalization is done to reduce the effect of redundancy of the pixel values. To incorporate these SDPs into a feature vector, the mean of these SDPs are computed using Equation 4.5, which is defined as:

$$\mu_X = \frac{1}{M}\sum_{k=1}^{m} H_X(k), \mu_Y = \frac{1}{N}\sum_{l=1}^{n} V_Y(l) \tag{4.5}$$

where μ_X and μ_Y are the mean in x and y directions respectively. Hence the feature vector is represented as:

$$f_{SDPs} = [\mu_x, \mu_y]. \tag{4.6}$$

In Figure 4.5, the SDPs of different human activities are shown. The variation of SDPs of different activities are true reflection of the shape change in the horizontal and vertical directions. In the first row of Figure 4.5, the AESI of bending activity is represented in terms of SDPs of x and y directions, the variation of SDPs in x-direction is initially less because the upper part of human body is in minimal motion and thus reflect less spatio-temporal change as compared to the middle part of body, that’s why peak is reflected in the middle of SDPs of x-direction. Similarly, the representation of SDPs in y-direction can be interpreted. SDPs of all the activities represented in Figure 4.5 shows the true reflection of AESI. It is also very important that every plot should be different because activities are different, hence it can be said that SDPs are the discriminative features.

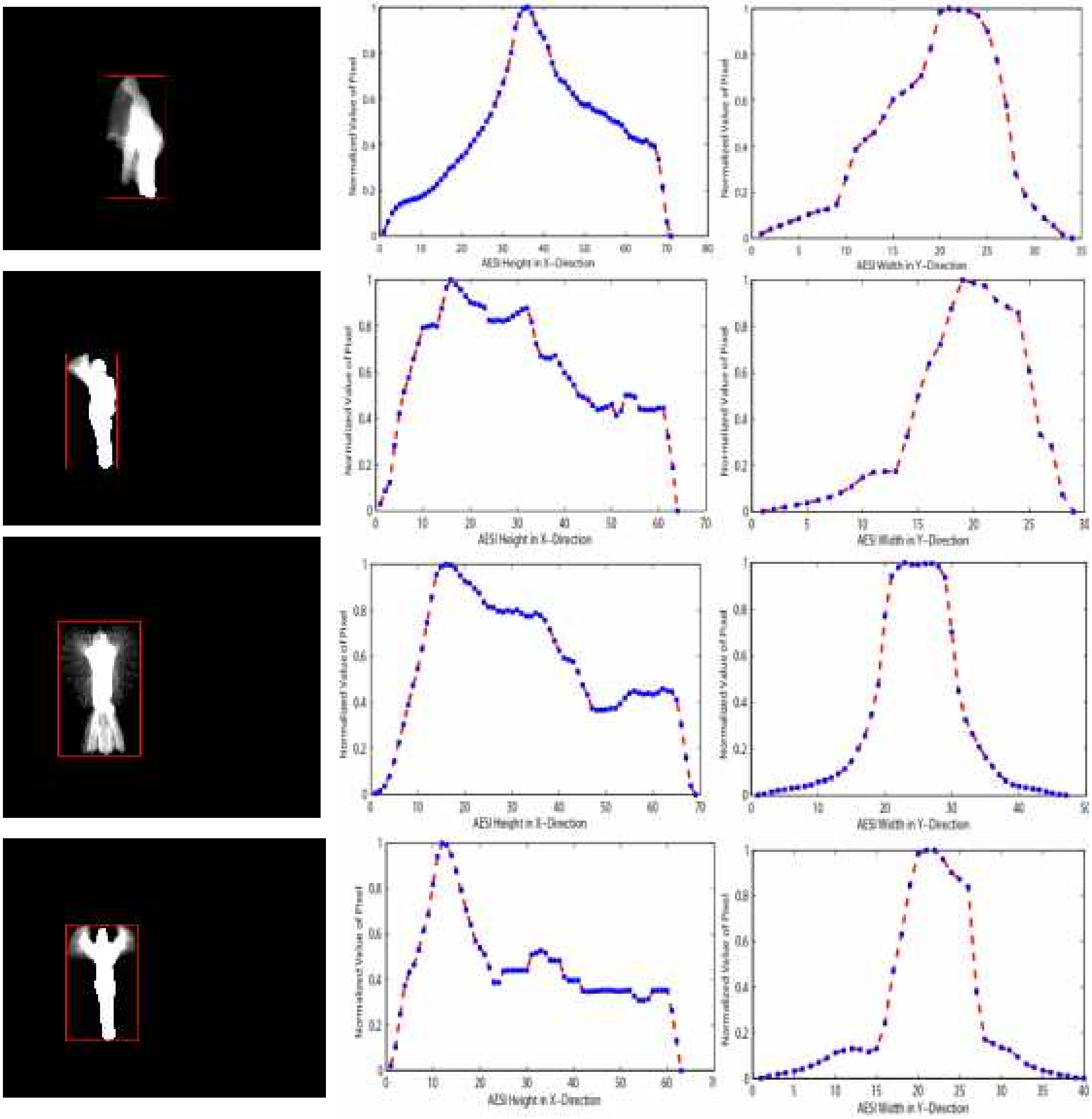


Figure 4.5: SDPs represenation of different human activities.

Up to this stage, the shape of human activity is represented by using SDGs and SDPs but the problem with these features is that when the activities have more similarity to each other like walking and running then these features reflect the feature values, which may be close to each other. Hence, in general for the recognition of human activity shape alone can't give sufficient information, hence temporal information is also needed for the representation of all the activities. Therefore, in the next section a motion based temporal information is computed through □-Transform and at recognition stage it is merged with shape information.

4.2.2 Computation of Rotation using $\mathfrak{R}$-Transform

The rotation of silhouettes gives the directional as well as the temporal motion information of human body and these are computed using □-transform. There are few approaches [111] [112] [113] [114], where $\mathfrak{R}$- Transform is used to estimate the shape of rotating objects as well as rotation of the 2D human postures [115]. Initially, Tabbone et al. [116] defined □-transform by using Radon Transform (R_T) of the binary silhouettes of the human activity. Consider a sequence of silhouette image $I_t(x,y)$ of the human activity, where t is the frame number and subsequently □-Transform □(θ) is defined as follows:

$$□(\theta) = \int_{-\infty}^{\infty} R_T^2\,(\rho,\theta)\partial\rho \tag{4.7}$$

where R_T is the Radon Transform, which gives the directional features in the range of angle$(0 - 179^0$) and is defined as the integral of a continuous image $I(x,y)$ from $-\infty$ $to\ \infty.$

$$R_T(\rho,\theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(x,y)\delta(\rho - x\cos\theta - \sin\theta)\,\partial x\,\partial y \tag{4.8}$$

where $\delta(.)$ is defined as the Dirac delta function which is zero everywhere except at the origin, where it approaches infinity. The perpendicular distance ρ from the origin to the radon line is calculated using Equation 4.9 and shown in Figure 4.6 (a, b).

$$\rho = xcos\theta + ysin\theta \quad for\ (0 \le \theta \le \pi), (-\infty \le \rho \le \infty) \tag{4.9}$$

where 'θ'is the angle between horizontal axis and the projection line.

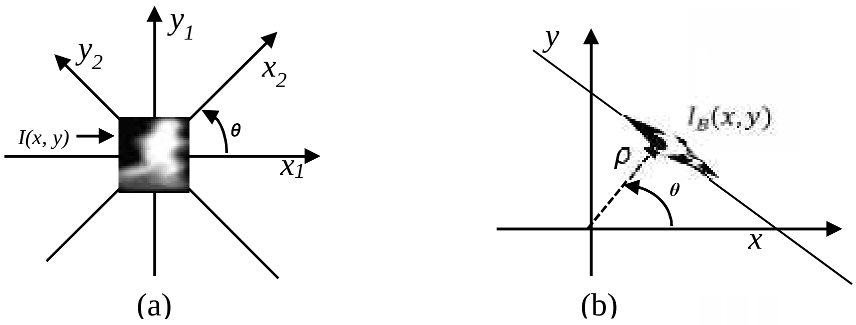


Figure 4.6: Depiction of projection lines over 2D image $I(x, y)$.

$I_B(x, y)$ is the binary image, which is defined as:

$$I_B(x, y) = \begin{cases} 1 & if\ (x, y) \in F \\ 0 & otherwise \end{cases} \tag{4.10}$$

where F is region of binary silhouette. Radon transform cannot restore all the parameters of the original geometric transformation when the image is translated or rotated or scaled. Hence, [116] introduces the □- transform, which is invariant to translation and scaling parameters but sensitive to the rotation, hence it provides sufficient discriminative changes for orientation for various shapes. The normalized □-transform ($\square_{\text{norm}}(\theta)$) improves the similarity measure and compactness of feature representation, which is defined as:

$$\square_{norm}(\theta) = \frac{\int_{-\infty}^{\infty} \Re(\theta) d\theta}{max(\Re(\theta))} \tag{4.11}$$

The fundamental properties of □-transform show that it is invariant against the scaling and translation but sensitive to the rotation and these properties can be proven theoretically as well as experimentally. To verify these properties of $\Re$-transform, let us consider that the original image is scaled by a translated by (x', y') and rotated by θ_0. For scaling of $\Re$-transform, it can be expressed as:

$$\frac{1}{a^2}\int_{-\infty}^{\infty} R_T^2(a\rho,\theta)d\rho = \frac{1}{a^3}\int_{-\infty}^{\infty} R_T^2(m,\theta)\,dm = \frac{1}{a^3}\Re(\theta) \tag{4.12}$$

From Equation 4.12, it can be seen that □-transform is independent of scaling, a similar thing can be seen from the Figure 4.7, which is verified using MATLAB. Similarly it can be expressed for translation and rotation of $\Re$-transform, which are as follows:

$$\int_{-\infty}^{\infty} R_T^2((\rho - x'\cos(\theta) - y'\sin(\theta)),\theta)d\rho = \int_{-\infty}^{\infty} R_T^2(m,\theta)\,dm = □(\theta) \tag{4.13}$$

$$\int_{-\infty}^{\infty} R_T^2(\rho,(\theta+\theta_0))d\rho = \Re(\theta+\theta_0) \tag{4.14}$$

Equation 4.13, and 4.14 respectively shows the invariance to translation and variance to rotation of □-transform. A similar verification of these properties can be seen in Figure 4.8, where the Radon Transform and □-transform of a silhouette frame of running activity is shown with the translation, scaling and rotation.

In Figure 4.7, rotation in the image shows more variation in the brighter portion of R_T when compared to other images because in the rotation, there is more deviation in the pixel values corresponding to projection lines. The magnitude of the translated image varies as compared to the scaled image, but the signal representation of □-transform remains the same. The rotational sensitivity of □-transform is used for the representation of motion temporal information of different activities and this approach is more effective if activity is having more rotation than translation like abnormal activities (vomiting, forward fall, backward fall, etc.) [117]. Hence, from Figure 4.7 the robustness of □-transform is proven against the translation and scaling, which is very essential property of feature vector.

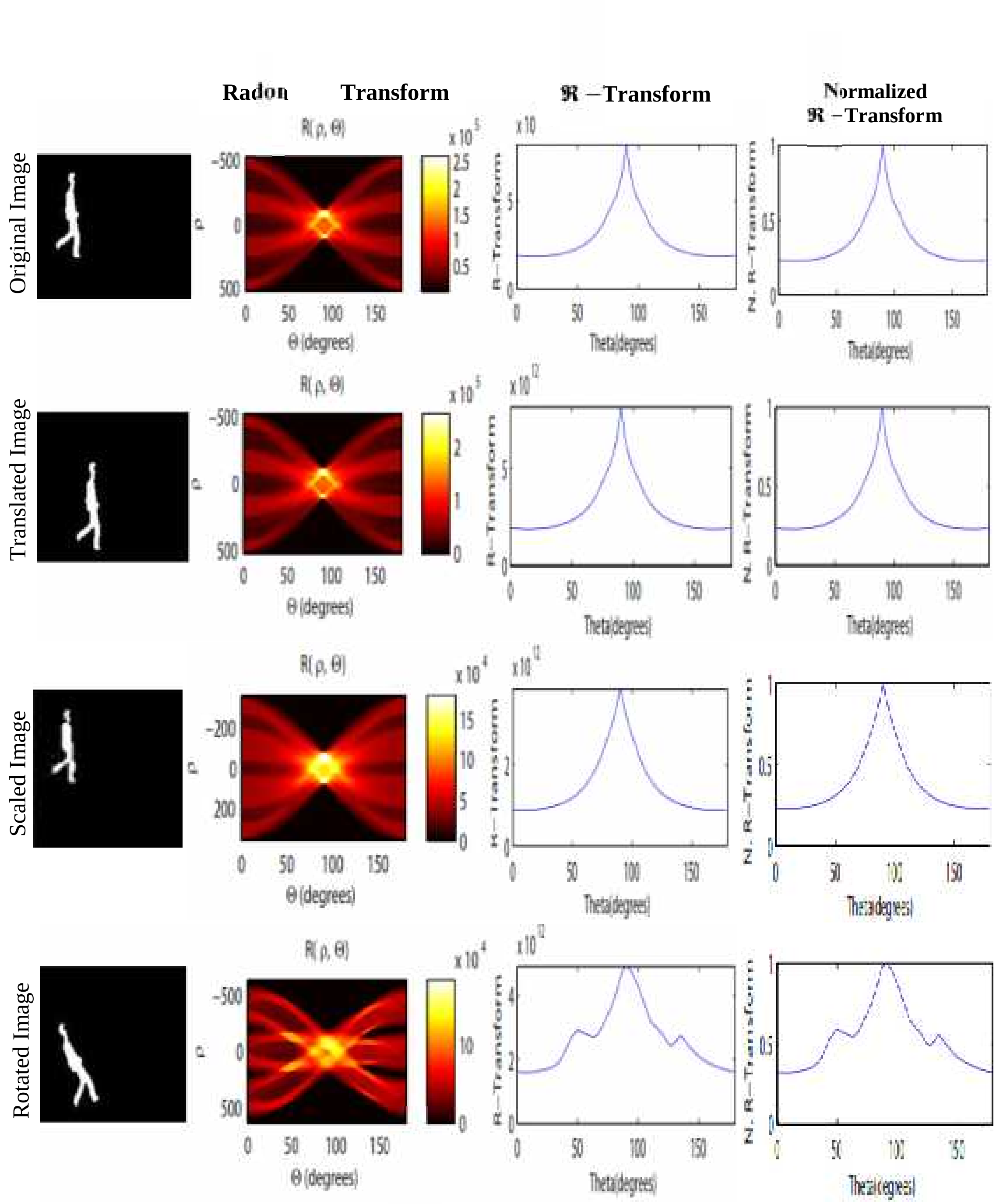
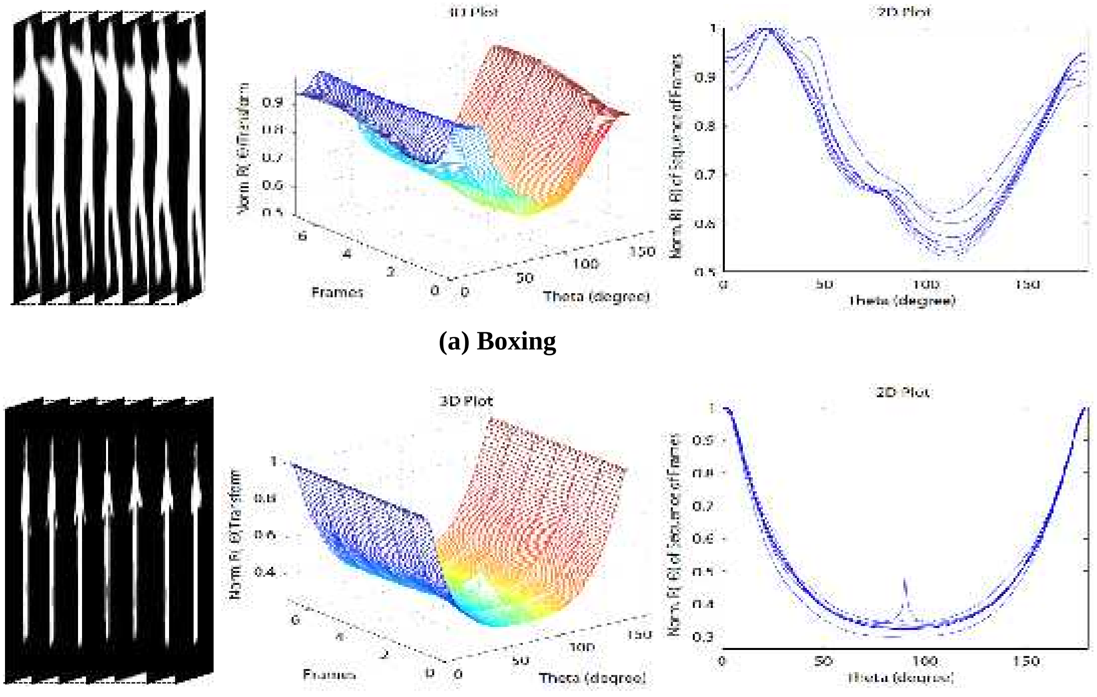


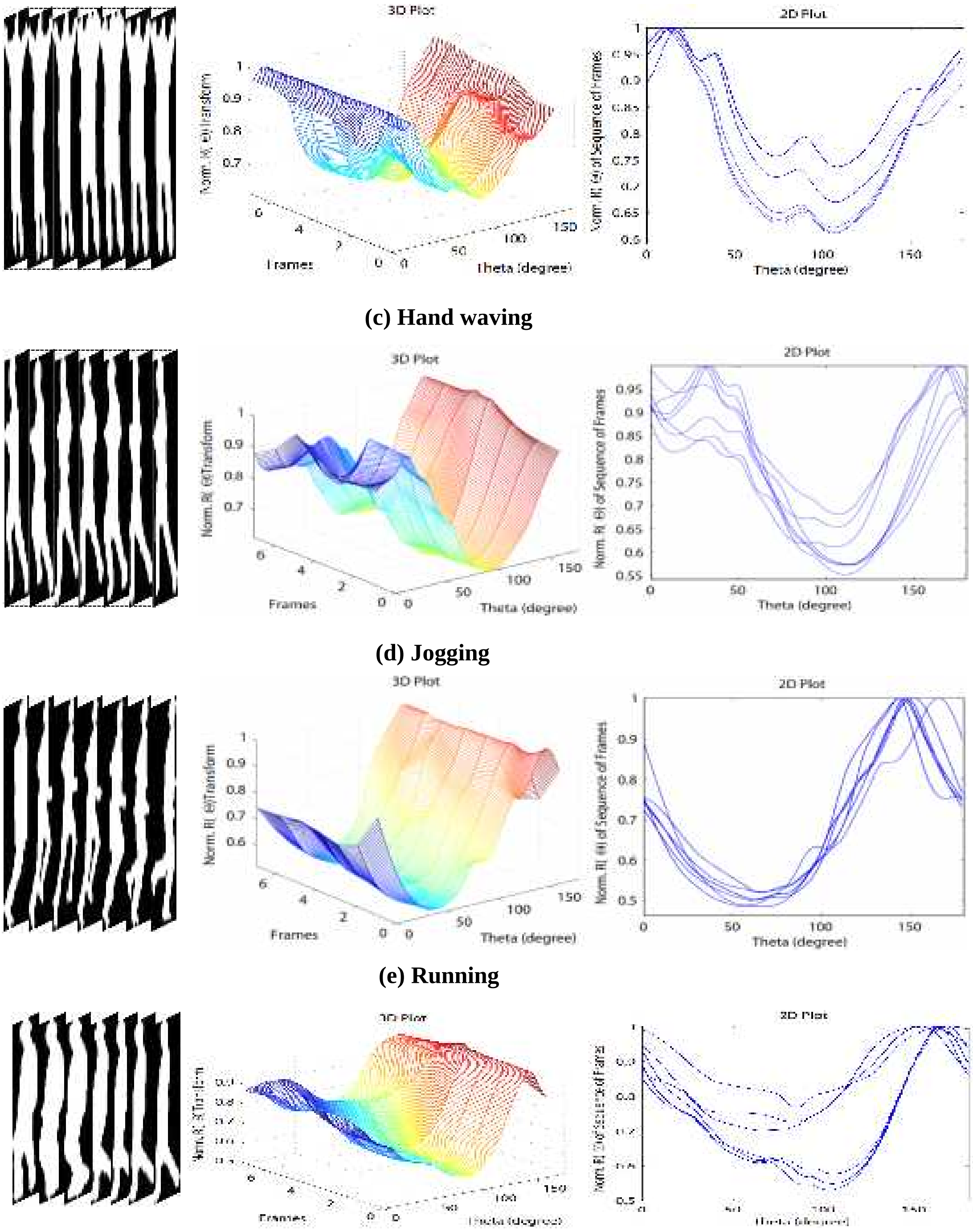
Figure 4.7: Depiction of ℜ-transform robustness and sensitivity.

The □-transform representation of different activities of a video sequence is shown in Figure 4.8. The activities presented in the Figure 4.8 are boxing, handclapping, hand waving, jogging, running and walking. The first column of Figure 4.8 is the human silhouette video sequence, second column is the 3D plot of normalized □-transform, and third column is the 2D plot of normalized □-transform. From the representation of □-transform, it is observed that the □-transform of different types of activities is significantly different. The geometrical profiles of normalized □-transform for jogging and walking actions look quite similar due to the true nature of these activities. Hence, it can be inferred that □-transform representation is proficient in describing the motion characteristics of the human action but alone it cannot sufficiently provide the information for distinguishing the activities. To take out the motion temporal information from □-transform, a group of frames are chosen and their □-transform are evaluated, which are further used to integrate with SDGs and SDPs. The dimension of □-transform of a $M \times N$ silhouette image is 1×180 and for k frames it is $k \times 180$, which is a high dimensional feature hence the dimension of this feature must be reduced by using proper dimensional reduction techniques.



(a) Boxing

(b) Hand clapping



(c) Hand waving

(d) Jogging

(e) Running

(f) Walking

Figure 4.8: Representation of ℜ-transform for different activities.

Principal Component Analysis (PCA) is one of the most popular and simplest technique used for dimension reduction of data, the reduction of dimension of data is done by keeping the maximum variability of the data and ignoring the less variant data. The operation of PCA [81] reveals the internal structure of the human silhouette in a manner that proficiently explains the variance in the data. The dominant features of dataset are obtained by solving the Eigen value problem of Covariance matrix of the dataset, represented as: The dimension of $[k \times 180]$ feature vectors are reduced to discriminative feature vector of size $[1 \times k]$ with the help of PCA.

4.2.3 Final Feature Formation Model

The final feature representation of an activity sequence is done by concatenation of the computed features. The detailed flow of final feature formation model is as shown in Figure 4.9.

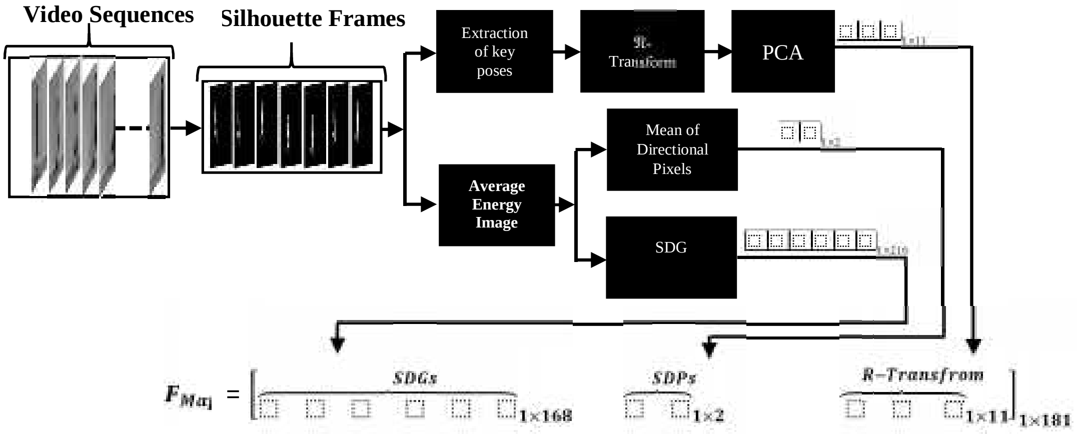


Figure 4.9: Flow diagram of final feature formation model.

The details of numerical figure given in the Figure 4.9 is based on the experimental parameters used, the value of $K = 8$ and $= 2$. The shape based SDGs and SDPs features are combined with the rotational feature computed using the □-transform. The

SDGs and SDPs have a dimension of 1×168 and 1×2 respectively. Further, these feature vectors are combined with □-transform features computed on the binary silhouettes. The □-transform computed feature vector has the dimension of 11×180. Here, the number of key frames used are 11. The dimension of □-transform feature is reduced using PCA, which gives the reduced discriminative feature vector of 1×11. Finally, these features are concatenated and give the resultant feature vector of $[1 \times 168 + 1 \times 2 + 1 \times 11] = [1 \times 181]$ dimension. The performance of final descriptor is evaluated by conducting an experiment on the standard dataset.

4.2.4 Experimental Results, Comparison and Analysis

To evaluate the effectiveness of the proposed framework, experiments are conducted using publically available standard datasets of Weizmann and KTH. The recognition accuracy of the proposed methods is computed using SVM and LOOCV as a test scheme. The detail about these datasets and the process of computing recognition accuracies are explained in chapter 3.

4.2.4.1 Results of Weizmann Dataset

The extraction of binary silhouette on this dataset is very accurate and effective due to moderate change in the recording setting and less intra-class variability and clean background. The explanation of Weizmann dataset set is given in the experimental section of Chapter 3. The SDGs and SDPs computation on this dataset is very effective and highly discriminative among various activities. The □-Transform signals representations are also very smooth and variant. Hence, the average recognition accuracy (ARA) computed on this dataset is almost 100%. The other reason for the high accuracy on this dataset is the limitations of this dataset itself. The dataset has enough number of classes with significant amount of interclass similarity but the number of test sample of the video is less. Hence, even for single misclassified sample, it leads to high drop in the recognition accuracy and if instead the sample is correctly classified then accuracy increases. Therefore, to test the algorithm on more challenging aspects of the dataset in terms of recording setting and adequate number of samples of videos of each classes of the activity, another experiment is done on KTH dataset. The description of KTH dataset is given in the experimental section of Chapter 3. The ARA achieved on this dataset is compared with the similar state-of-the-art methods and demonstrate the relatively higher recognition accuracy with earlier techniques.

4.2.4.2 Results of KTH Dataset

This dataset is more challenging in terms of intra-class variability, illumination change, camera movement etc. The ARA achieved on this dataset for each class of the activities are as shown in Table 4.1. It has two similar classes of activities i.e. jogging and running which significantly decrease the accuracy as compared to other activities. Moreover, the extraction of human silhouette is difficult for this dataset due to the variation of lighting conditions, which may cause the reduction of accuracy. The average recognition accuracy (ARA) achieved on this dataset is 95.50%. The ARA achieved on this dataset is compared with the techniques of others on similar state-of-the-art datasets [60] [118] [119] [120] [121] and gives superior result as depicted in Table 4.2.

Table 4.1: Classification results on KTH dataset using SDGs, SDPs, ℜ-Transform.

Activities\Classifier	Boxing	Hand waving	Hand clapping	Jogging	Running	Walking
SVM (%)	100	100	92	92	89	100

Table 4.2: State-of-the-art comparison with the techniques of others.

Techniques	Weizmann Dataset	KTH Dataset
Dollar et al. [60]	82.50	81.17
Nibeles el al. [118]	90.00	83.33
Kläser el al. [119]	84.30	91.40
Hung et al. [120]	92.50	35.04
Benmoktar [121]	89.00	92.5
Proposed approach	**100**	**95.50**

It is also apparent that the SDGs play an important role in the recognition and the better spatial distribution around the ROI further improves the recognition accuracy. The effect of SDPs are not significantly higher as compared to SDGs and □-Transform.

In this experiment, it is seen that the computation SDGs features is dependent on the number of orientation bins and number of decomposition levels and to extract out accurate spatial distribution of an image, it is very important to select these parameters appropriately. Mostly it is seen that the high value of orientation bins and number of decomposition levels gives better spatial distribution but the dimensionality of feature vector increases exponentially. Hence, to select appropriate value of these parameters, one has to optimize the values of these parameters or to change the ways of computation so that better spatial distribution can be achieved without much increase in the dimensionality. To keep the concerns of dimensionality as well the quality of spatial distribution, an alternate way of computation of spatial distribution is described in the subsequent section, which is based on the analysis of spatial distribution at various decomposition levels.

4.3 Modified SDGs approach

This approach is an extension of the previous model of SDGs computation. The main aim of this approach is to provide an improved spatial distribution computation model with minimum cost of computation. As it is seen that the SDGs computation at K orientation bins and v decomposition level, the feature vector is $K\sum_v 4^v$. To get the better estimate about the SDGs feature vector size and its effect, let us consider an example where $K = 8$ and $v = 0$,1,2,3

At $v = 0$, $SDGs_{v=0} = 8\sum_v 4^0 = 1 \times 8$ (4.15)

At $v = 1$, $SDGs_{v=1} = 8\sum_v 4^v = 8[1 + 4] = 1 \times 40$ (4.16)

At $v = 2$, $SDGs_{v=2} = 8\sum_v 4^v = 8[1 + 4 + 16] = 1 \times 168$ (4.17)

At $v = 3$, $SDGs_{v=3} = 8\sum_v 4^v = 8[1 + 4 + 16 + 64] = 1 \times 680$ (4.18)

From Equation 4.15-4.18, it can be seen that the size of SDGs feature vector increases exponentially but effect of SDGs feature vector at higher level of decomposition is compared (level 2 and 3). The magnitude of spatial distribution computation is decreased as there is an increase in the level of decomposition which can been seen in Figure 4.11. Hence, it can be said that to get better and more effective spatial distribution, the higher level of decomposition isn’t be a good approach in respect to effect as well as dimension. Therefore, to enhance the effect of SDGs with less computation as compared to level 3, a modified approach of computation of spatial

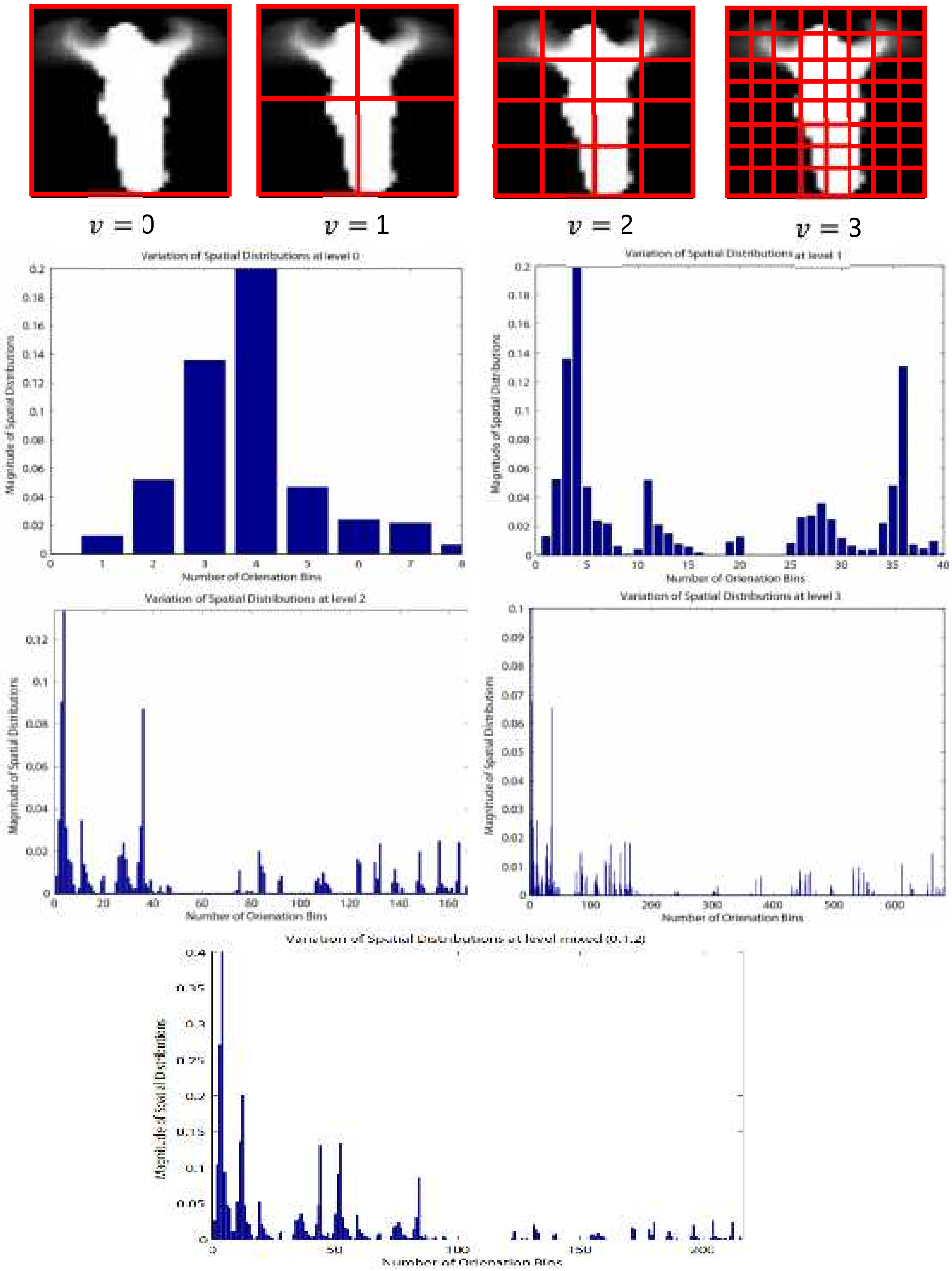


Figure 4.10: Representation of levels composition and spatial distributions.

Figure 4.10 shows the spatial distribution variation at various levels and highest level shown in the figure is 3, where the spatial distribution of an image shows that the magnitude values are decreasing very fast as compared to other levels. A mixed level of computation distribution is also present in the same figure, where the magnitude of spatial distribution is comparatively better than the level 3. The computation of spatial distribution of gradients at mixed level spatial distribution is as explained in algorithm 2.

```
Algorithm 2: Computation of modified SDGs
Input: Input an AESI $A_E(x,y)$ computed using Equation 4.1
Step1: Crop the ROI of AESI and denoted as $R_E(x,y)$ . To maintain uniformity,
        the dimension is resized to $M \times N$.
Step2: Compute the magnitude and gradient of each pixels of $R_E(x,y)$ image at
        different decomposition levels $(v)$ using Equation 4.2.
Step3: Quantize the computed magnitude of intensity at various levels $(v)$ into $K$-
        orientation bins.
Step4: Summing all the regions together at level $(v)$ $f_{SDGs} = K \sum_v 4^v$
Step5: Concatenate the level (0, 1, 2) of SDGs together and [illegible] modified SDGs
        feature vector is as: $f_{MSDGs} = K \left( \underbrace{\sum_v 4^v}_{level\ v=0} ; \underbrace{\sum_v 4^v}_{level\ v=1} ; \underbrace{\sum_v 4^v}_{level\ v=2} \right)$.
Output: The histogram of modified SDGs feature vector as shown in Figure 4.11.
```

The process of concatenation of previous levels of SDGs is carried out column-wise such as: $[SDGs_{level=0}; SDGs_{level=1}; SDGs_{level=2} \ldots; SDGs_{level=v}]$.

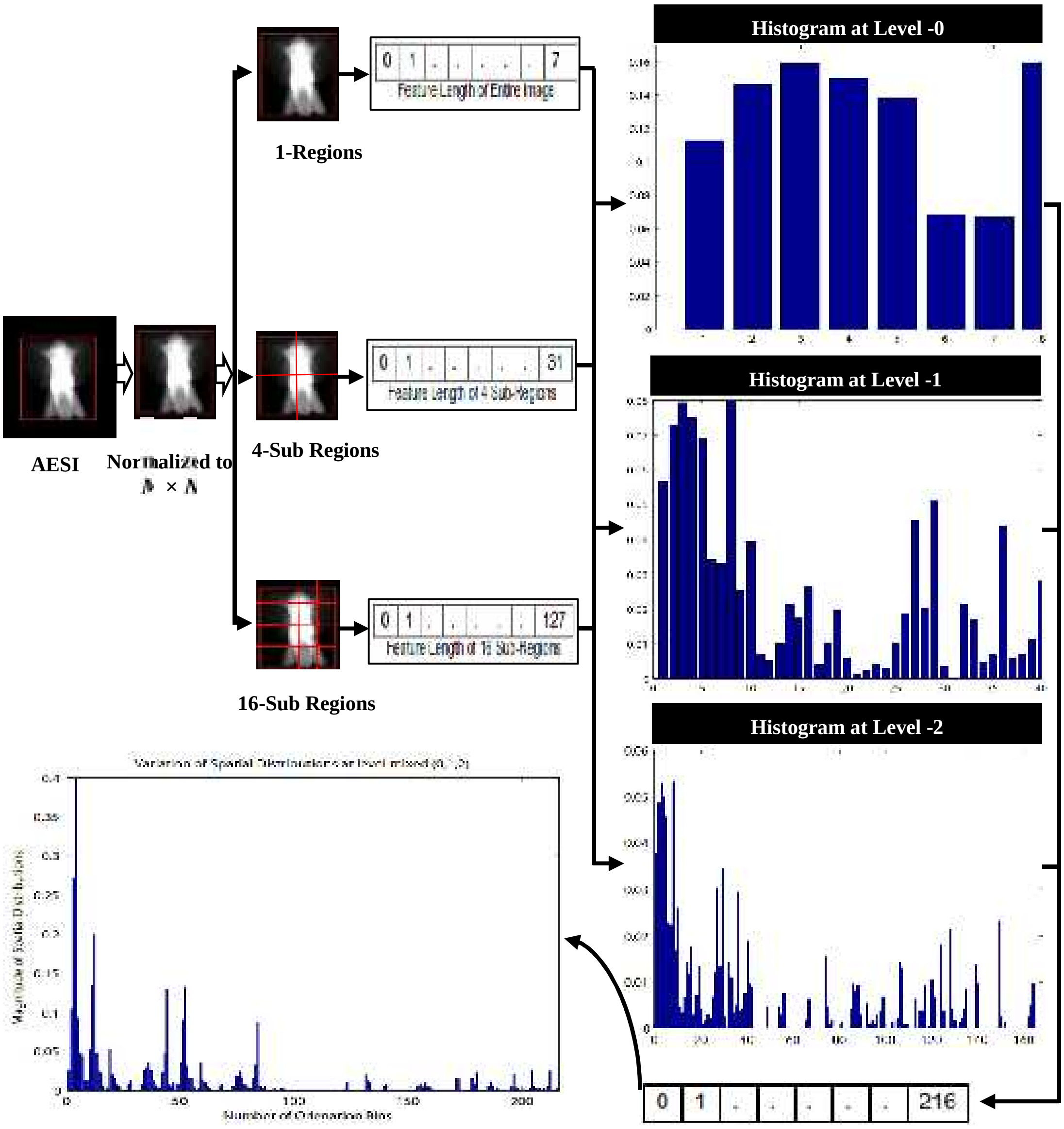


Figure 4.11: Depiction of modified SDGs computation at mixed level 'm' of 8-number of orientation bins.

Figure 4.11 shows the flow of histogram formation of SDGs at various levels of decomposition of AESIs of 8-orientation bins. The number of decompotion levels are 0, 1 & 2. Finally a mixed level 'm' histogram is also presented which has feature vector length 1 × 216.

4.3.1 Proposed Model for HAR using Modified SDGs

A HAR model is proposed by integration of modified SDGs (MSDGs) features with the □-transform and SDPs at the recognition stage. The flow diagram of proposed scheme is shown in Figure 4.12, where the input video sequence is the video feed obtained by CCTV/Video camera and the process of texture based segmentation is described in the Chapter 3.

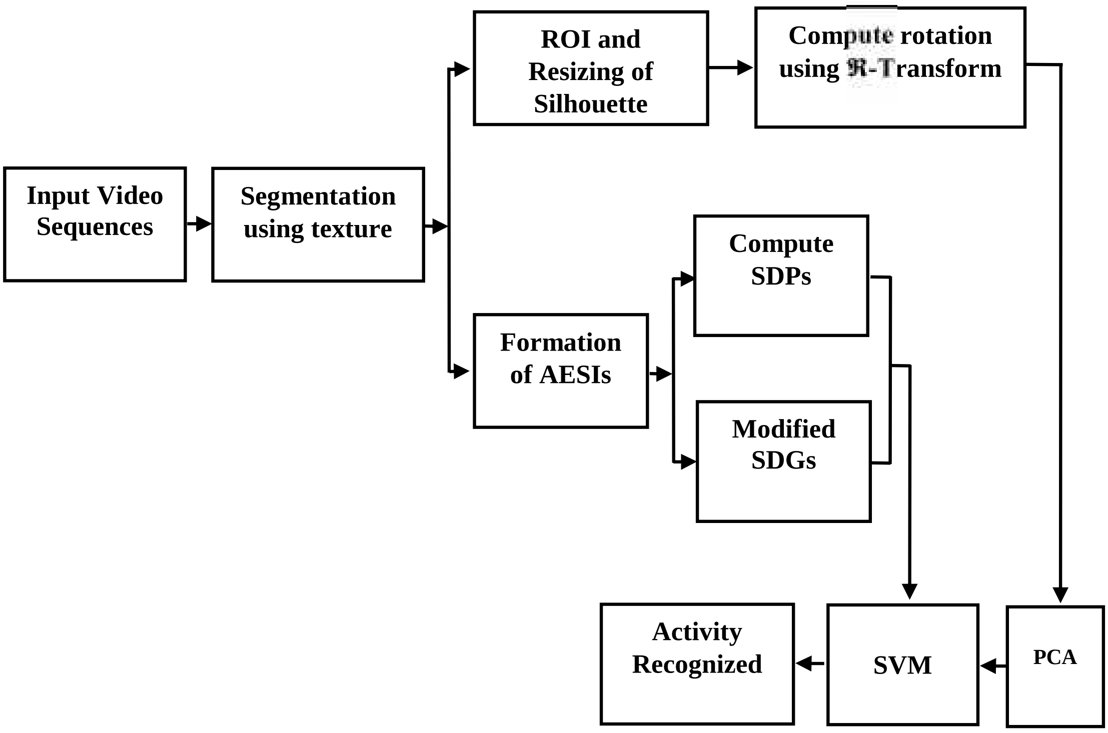

Input Video Sequences
Segmentation using texture
ROI and Resizing of Silhouette
Compute rotation using ℜ-Transform
Formation of AESIs
Compute SDPs
Modified SDGs
Activity Recognized
SVM
PCA

Figure 4.12: Work flow diagram of modified SDGs for HAR.

To get more localize information ROI is selected and further to maintain the uniformity resizing is done. The formation of AESIs, computation of □-Transform, and SDPs are as explained in the section 4.2. To know the effect of modified SDGs feature computation, an experiment is conducted at levels 0, 1, 2 and a mixed level (m) which is explained in the subsequent section.

4.3.2 Formation of Final Feature Concatenated Model

The final feature map model (F_{Map}) is formed by concatenating the MSDGs with the feature set obtained in section 4.2.1.2 and 4.2.2. The flow of concatenation of feature vectors is as shown in Figure 4.13. The shape based MSDGs feature vector and SDPs feature vector are concatenated with the rotational feature vector, which is computed using □-transform. This concatenation forms a new descriptor for the representation of human action and activity. This feature model is based upon the experimental settings used in this experiment.

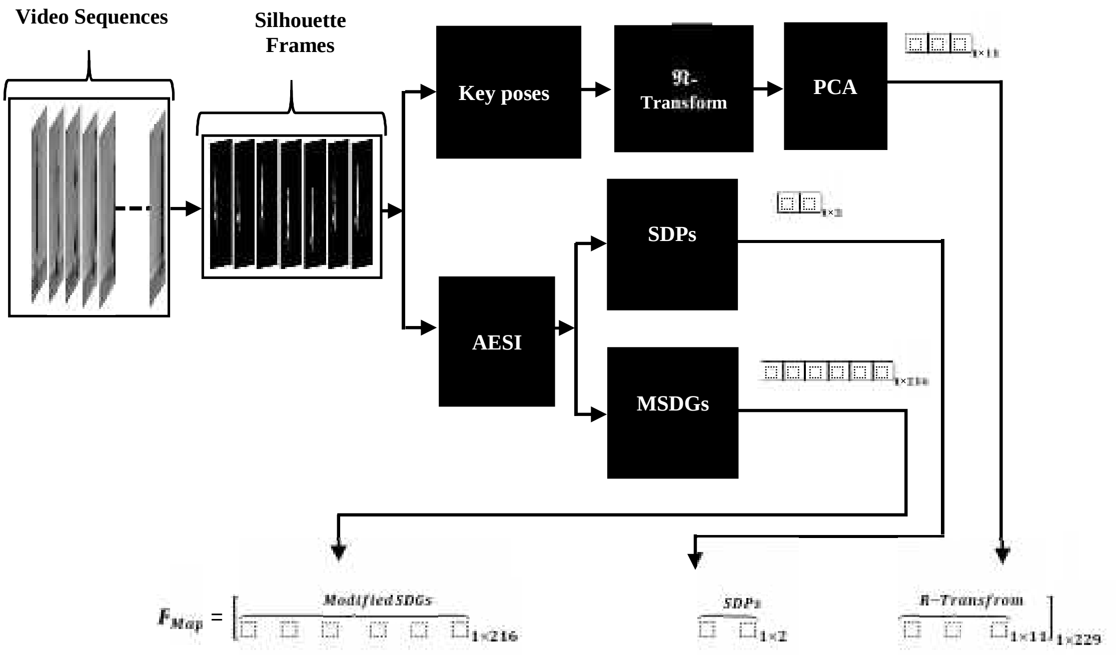


Figure 4.13: Flow diagram of concatenation of features.

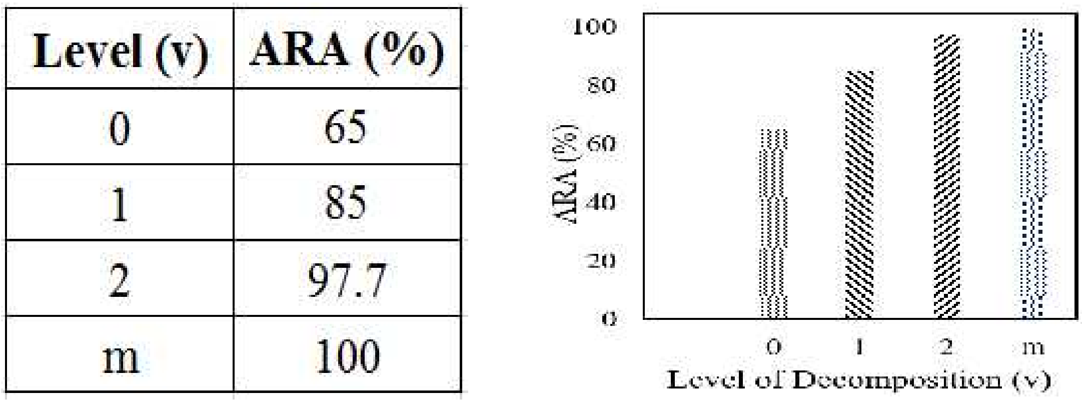
The MSDGs and SDPs feature vectors have dimension of 1×216 and 1×2 respectively. Further, these feature vectors are combined with □-transform features computed on the binary silhouette of size 50×50. The □-transform computed features have a dimension of 11×180. Here, the number of key frames used is 11. The dimension of □-transform feature is reduced using PCA, which gives the reduced discriminative feature vector of 1×11. Finally, the shape based MSDGs, SDPs feature vectors and motion based □-transform feature vectors are concatenated to produce the resultant feature map of $[1 \times 216 + 1 \times 2 + 1 \times 11] = [1 \times 229]$ size. The performance of final descriptor is evaluated through SVM using various standard datasets.

4.3.3 Experimental Results, Comparison and Analysis

In order to test the performance of the proposed approach, experiments are conducted using three publicly available datasets i.e. Weizmann [74], KTH [97], and Ballet [98]. The recognition accuracy is computed through SVM using leave-one-out cross validation strategy for all datasets. In leave-one out strategy, one sequence is used as a test video while others are used as training sets. This procedure continues until all the sequences are tested for a single activity and it further proceeds for other activities. The advantage of this method is that whole information of the dataset is used to calculate the accuracy. To know the effectiveness of proposed approach in comparison with the earlier techniques, a comparison table is presented for all the datasets at the similar state-of-the-art in terms of ARA. The classification strategies used in the different techniques are quite similar and are named as leave one out cross validation (LOOCV), leave-one-sequence-out cross validation (LOSOCV), leave-one-video-out cross validation (LOVOCV) and leave-one-person-out cross validation (LOPOCV). The classifiers used in the comparison are: SVM [77] [85] [58], KNN [72] [74] [99] and Fuzzy [122].

4.3.3.1 Recognition Results on Weizmann Dataset

In this dataset, the extraction of binary silhouettes does not offer great challenge due to the clean background and stable environmental settings. Hence, an accurate silhouette is extracted for all the videos of dataset, which depicts the effectiveness of the proposed approach. The detailed description of Weizmann human action dataset is given in the experimental section of Chapter 3. The result presented for this dataset gives the effectiveness of overall approach as well as the effect of SDG vector at different levels.



Level (v)	ARA (%)
0	65
1	85
2	97.7
m	100

Figure 4.14: ARA achieved on Weizmann dataset at different level of SDG.

From Figure 4.14, the recognition accuracy is increased with an increase in the level of computation of SDG vector. At level ‘0’ the value of ARA indicates the minimal effect of SDG vector, which results in ARA 65%, which is due to the dominance of □-

Transform and variation directional pixel feature vector. At the higher level (2, m) the ARA is 97.7 and 100% respectively, which indicates that the computation of spatial distribution is much finer at a higher level. The highest recognition accuracy is compared with the similar state-of-the-art and is shown in Table 4.3. The experimental setting used in this experiment is similar to the Gorelick et al. [74]. In all these methods, the inputs are the human silhouettes of activity. Hence, the comparison is relatively fair. The recognition accuracy achieved through proposed approach is higher than that of many earlier techniques.

Table 4.3: Comparison of ARA on Weizmann dataset.

Method	Classifier	Test Scheme	ARA (%)
Gorelick et al. [74]	KNN	LOOCV	97.5
Chaaraoui et al. [72]	KNN	LOSOCV	92.8
Wu & Shao [77]	SVM	LOSOCV	97.78
Goudelis et al. [85]	SVM	LOPOCV	95.42
Melfi et al. [58]	SVM	LOOCV	99.02
Touati & Mignotte [99]	KNN	LOOCV	92.3
Yao et al. [122]	FUZZY	LOOCV	94.03
Proposed Method	**SVM**	**LOOCV**	**100**

4.3.3.2 Recognition Results on KTH Dataset

This dataset is more challenging as compared to Weizmann dataset in terms of recording conditions, which includes zoom in and zoom out of the camera, variation in viewing angle, and lighting conditions. The exhaustive introduction of this dataset is given in the experimental section of Chapter 3. Due to these variations, the extraction of silhouette is difficult as compared to the previous dataset. Therefore, the less ARA is achieved. As it is seen in Figure 4.14, the highest ARA is achieved at level 2 and ‘m’. Hence, for this dataset the ARA is computed at the level of 2 and ‘m’. The ARA achieved on this dataset is as presented in Table 4.4.

Table 4.4: Recognition results on KTH dataset using MSDGs, SDPs, & ℜ-Transform.

Activity \ SDGs Level	**Boxing**	**Hand-waving**	**Hand-clapping**	**Jogging**	**Running**	**Walking**	**ARA (%)**
2	100	100	92.5	92.5	90	100	95.83
‘m’	100	100	95	95	92.5	100	97.08

The highest ARA achieved on this dataset is 97.08% which is higher than the ARA achieved at the level of 2. Moreover, from Table 4.4, the recognition accuracy of the similar classes like Jogging, and Running at the ‘m’ level is boosted. Hence, it can be said that the ‘m’ level is more effective than the individual level of SDG computation. The highest ARA achieved is compared with the similar state-of-the-art techniques and presented in Table 4.5.

Table 4.5: Comparison of ARA on KTH dataset.

Method	**Input**	**Classifier**	**Test scheme**	**ARR (%)**
Saghafi & Rajan [100]	Silhouettes	KNN	LOOCV	92.6
Goudelis et al. [85]	Silhouettes	SVM	LOPOCV	93.14
Melfi et al. [58]	Silhouettes	SVM	LOOCV	95.25
Rahman et al. [101]	Silhouettes	KNN	LOOCV	94.49
Benmokhtar [121]	Images	SVM	LOOCV	92.50
Vishwakarma & Kapoor [29]	Silhouettes	SVM-NN	LOOCV	96.4
Proposed Method	**Silhouettes**	**SVM**	**LOOCV**	**97.08**

4.3.3.3 Recognition Results on Ballet Dataset

The recognition accuracy is computed on this dataset at level ‘m’ because it is seen from earlier experiment that the mixed ‘m’ gives a higher recognition accuracy as compared to other levels. The details of classification result of different activities of this dataset is presented in Table 4.6. The highest ARA achieved is 94.5%, which is less than the earlier datasets used in this work. The key reason for the decrease in the accuracy is the complex motion pattern, which is different for execution of the motion from actor to actor and high misclassification error is introduced due to the “hopping” movement as it is confused with a much related movement “jumping”. It is also observed that the formation of the AESIs is difficult because of the complexity. Where b1, b2, b3, b4, b6, b7, and b8 are hopping, jumping, left-to-right hand opening, leg swinging, right-to-left hand opening, standing hand opening, standing still and turning ballet movement respectively. The detailed explanation about the Ballet movement dataset is described in the experimental section of Chapter 3.

Table 4.6: Recognition accuracy (%) on Ballet dataset using MSDGs, SDPs, & ℜ-Transform.

Activity \ SDG Level	b1	b2	b3	b4	b5	b6	b7	b8	ARA
‘m’	90	72	100	100	100	100	100	94	94.5

Table 4.7: Comparison of ARA on Ballet dataset.

Method	Input	Classifier	Test scheme	ARA (%)
Fathi & Mori [98]	Images	Adaboost	LOOCV	51
Wang & Mori [103]	Images	S-CTM	LOO	91.3
Guha & Ward [104]	Silhouettes	RSR	LOO	91.1
Ming et al. [105]	Silhouettes	RVM	LOO	90.8
Iosifidis et al. [106]	Images	SVM	LOO	91.1
Vishwakarma & Kapoor [29]	Silhouettes	SVM-NN	LOOCV	94.0
Proposed Method	**Silhouettes**	**SVM**	**LOOCV**	**94.5**

The ARR achieved on this dataset is compared with original work [98] on this dataset as well as the work of others on similar state-of-the-art and presented in Table 4.7. The experimental settings used for this dataset are similar to those in the original paper. The performance of proposed approach can be clearly seen from the comparison Table 4.7, where the recognition rate achieved through this approach is comparatively fair and performance is significantly improved.

4.4 Significant Findings

The in-depth analysis of the SDG feature vector with the integration of □-transfrom and directional variation of pixels reveals promising annotations:

- The recognition accuracy achieved with this approach is better than that of the many earlier existing approaches due to the additional fusion of motion information in AESIs. Hence, the proposed approach has a good discriminative power of representing different kinds of activities.
- The final feature descriptor formed with the integration of mixed level 'm' of SDG, □-transfrom and directional variation of pixels is superior in terms of recognition accuracy as compared to the individual level (0, 1, 2) of SDG features integration.
- Significant improvement in the recognition accuracy on KTH and Ballet dataset shows the robustness of proposed approach against the variation of lighting conditions, intra class dissimilarity of actions in terms of speed, and spatiotemporal scaling, the camera zooming in and zooming out, clothing, etc.
- As the number of key poses increase for the estimation of motion information using □-Transform, the dimension of the feature vector increases and the speed of computation also drops significantly.

- For computation of shape using SDGs, an increase in the number of levels of bins gives improved accuracy in the recognition rate but the dimension of the distribution vector is increased rapidly, which makes the system slow.
- The effect of directional pixel’s mean value on the recognition accuracy is not so high as compared to SDGs and □-Transform.

This chapter is based on the following work:

D.K. Vishwakarma, Rajiv Kapoor, "Integrated Approach for Human Action Recognition using Edge Spatial Distribution, Direction Pixel, and R -Transform", ***Advanced Robotics***, (**Publisher: Taylor & Francis**), 2015, Vol. 29, No. 23, pp. 1551-1561. [123]

CHAPTER 5

CONCLUSIONS & FUTURE SCOPE

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

5.1 Conclusions

The two major approaches of the human activity recognition based on human silhouettes are presented and these approaches are as follows:

- In the first approach the activity recognition is done by selecting the key poses of human silhouette and further to represent these silhouettes as a scheme of grids and cells. The problem of low recognition rate under the variant environmental conditions has been addressed by employing: (a) Accurate human silhouette extraction through texture based background subtraction approach; (b) Simple and effective representation of human silhouettes by means of grids and cells. (c) An effective hybrid classification model of "SVM-NN". The effectiveness of the proposed approach is tested on three publicly available datasets through LDA, KNN, SVM, and "SVM-NN" classification models and ARA of these models are measured. The success of these classification models is assessed using MCE and it is observed that the hybrid classification model i.e. the "SVM-NN" gives the least classification error. The overall performance of the proposed approach is found to be comparatively more effective. The parameters used for feature representation are simple and easily computable. The three datasets used here vary in terms of lighting conditions, indoor and outdoor environment, zoom in, zoom out, and hence it can be concluded that the proposed approach is robust under such diverse conditions. Despite the satisfactory results, some concerns have cropped up: (i) It is imperative that only one person is in the video sequence, (ii) Some parameters like the number of key poses, size of grids and cells can be further optimized (iii) This approach is less effective, when object is fully occluded.

- The second approach of human activity recognition is based on shape and motion features of the human silhouette in the video sequence, which enhances the recognition accuracy under challenging environmental conditions and complex motion pattern using average energy silhouette images and □-Transform. The shape of AESIs are computed using SDGs, MSDGs, and SDPs. The orientation based motion flow of human silhouette is computed using □-Transform. As the number of decomposition levels increases in the SDG vector, better quality of edge distribution and accurate results are achieved, but the complexity of the system increases due to increase in the vector size. The characteristics of edge distribution is also dependent on the ROI. The □-Transform computed feature vector results in high dimension and hence a dimension reduction technique is applied for the compact representation and improved classification. This unified technique supplies numerous distinctive feature vectors, which escort us to a robust and noise free action recognition model.

5.2 Future Research Scope

Despite the satisfactory results, some experimentation related issues are reported and these issues may lead to the future research paradigms. The issues are the computation time needed by these approaches, the dimension of feature vector, segmentation of silhouette, number of key parameters used and variability of the activity sequence (intra-class dissimilarities and inter-class similarities). To address the problem of high computation time the various check points used in the algorithms may be eliminated by using proper analysis so that the computation required may be reduced. Due to high dimensionality the classification is slow, hence, proper dimensionality reduction techniques may be utilized by keeping the maximum variability of features. Another approach, which may provide fast recognition of human activity is by using visual information based still images. The extraction of human silhouette under the cluttered background, illumination change, and camera motion may be carried out more accurately. There are various parameters used in these approaches like the size of cells, number key frames, number of orientation bins and level of decomposition which may be optimized to get higher recognition accuracy. To deal better with inter-class similarity and intra-class dissimilarity, a more robust multi classifier system may be designed so that misclassification may be reduced. In this study it is imperative that only one person is present in the video sequence. When object is occluded the approaches tends to be less effective. Hence one may use multiple camera based information to resolve the occlusion. The same approaches may be extended for other avenues of research like Human Style Recognition, Hand Gesture Recognition, Facial Recognition etc.

5.3 Future Applications

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

- Automatic surveillance need is increasing by the day for monitoring the public places like shopping malls [11], railways stations, bus stand, parking area etc. Due to heavy traffic, it is becoming difficult to monitor the data collected from various sensors manually. Hence, an automatic surveillance system is needed which can monitor the event occurring is sequence and raise an alarm when an unusual activity occurs.
- The system may be used as a health care system [124] [125] [126] for elderly, where the abnormal activities can be detected. Most of the office going people are away from their homes while their aged parents stay back at home. At this age, the chances of abnormal activities occurring are more substantial [127], like vomiting, falling from the floor [128], headache etc. Hence to provide assistance at the appropriate time, this system may come out to be very useful.
- A family doctor can measure the level of disorder of Gait [129] [130]of a person affected by certain diseases like the Parkinson, Elephantiasis, Poliomyelitis etc. With reference to the normal Gait of a person, a Gait disorder due to disease can be measured and accordingly prescription may be given.
- A system which can coach athletes to improve their techniques by providing correct assistance e.g. golf swing, cricket swing etc.

- A system which can help athlete by improving their projectile throw for example, disc throw, hammer throw, javelin throw, archery etc. The projectiles may be effectively calculated to improve the performance of the athlete.
- A system which can monitor activities between the borders of two countries may also be setup, which can detect unusual activities or motion of foreign objects.
- An intrusion based system can be developed for the purpose of security at home or offices.

REFERENCES

[1] J. Agrawal and M. Rayoo, "Human Activity Analysis: A Review," *ACM Computing Survey,* vol. 43, no. 3, pp. 16-43, 2011.

[2] T. D. Orazio and M. Leo, “A review of vision-based systems for soccer video analysis,” *Pattern Recognition,* vol. 43, no. 8, pp. 2911-2926, 2010.

[3] R. Poppe, “A survey on vision-based human action recognition,” *Image and Vision Computing,* vol. 28, no. 6, pp. 976-990, 2010.

[4] P. Ochs, J. Malik and T. Brox, “Segmentation of Moving Objects by Long Term Video Analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 36, no. 6, pp. 1187-1200, 2014.

[5] Y. J. Lee, J. Kim and K. Grauman, “Key-segments for video object segmentation,” in *IEEE International Conference on Computer Vision (ICCV)*, Barcelona, 2011.

[6] G. Zhang, J. Jia, W. Hua and H. Bao, “Robust Bilayer Segmentation and Motion/Depth Estimation with a Handheld Camera,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 33, no. 3, pp. 603-617, 2011.

[7] D. Ignakov, G. Liu and G. Okouneva, “Object segmentation in cluttered and visually complex environments,” *Autonomous Robots,* vol. 37, no. 2, pp. 111-135, 2014.

[8] M. Ollis and T. Williamson, “The future of 3D video,” *Computer,* vol. 34, no. 6, pp. 97-99, 2001.

[9] J. M. Batalla, , “Advanced multimedia service provisioning based on efficient interoperability of adaptive streaming protocol and high efficient video coding,” *Journal of Real-Time Image Processing,* Vols. 10.1007/s11554-015-0496-4, pp. 1-12, 2015.

[10] C. Deng, W. Lin, B.-s. Lee, C. T. Lau and M. T. Sun, “Performance analysis, parameter selection and extensions to H.264/AVC FRExt for high resolution video coding,” *Journal of Visual Communication and Image Representation,* vol. 22, no. 8, pp. 749-759, 2011.

[11] R. Arroyo, J. J. Yebes, L. M. Bergasa , I. G. Daza and J. Almazán, “Expert video-surveillance system for real-time detection of suspicious behaviors in shopping malls,” *Expert Systems with Applications,* vol. 42, no. 21, pp. 7991-8005, 2015.

[12] M. Valera and S. A. Velastin, “Intelligent distributed surveillance systems: a review,” *IEE Proceedings Vision Image and Signal Processing,* vol. 152, no. 2, pp. 192-204, 2005.

[13] L. Zhang, D. V. Kalashnikov, S. Mehrotra and R. Vaisenberg, “Context-based person identification framework for smart video surveillance,” *Machine Vision and Applications,* vol. 25, no. 7, pp. 1711-1725, 2014.

[14] R. Vezzani, M. Lombardi, A. Pieracci, P. Santinelli and R. Cucchiara, “A General-Purpose Sensing Floor Architecture for Human-Environment Interaction,” *ACM Transactions on Interactive Intelligent Systems,* vol. 5, no. 2, pp. 10-26, 2015.

[15] H. Hasan and S. A. Kareem, “Human–computer interaction using vision-based hand gesture recognition systems: a survey,” *Neural Computing and Applications,* vol. 25, no. 2, pp. 251-261, 2014.

[16] D. K. Vishwakarma and R. Kapoor, “An efficient interpretation of hand gestures to control smart interactive television,” *International Journal of Computational Vision and Robotics,* pp. 1-18, 2015.

[17] B. Ni, P. Moulin and S. Yan, “Pose Adaptive Motion Feature Pooling for Human Action Analysis,” *International Journal of Computer Vision,* vol. 111, no. 2, pp. 229-248, 2015.

[18] P. Turaga, R. Chellappa, V. S. Subrahmanian and O. Udrea, “Machine Recognition of Human Activities: A Survey,” *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 18, no. 11, pp. 1473 - 1488, 2008.

[19] A. A. Chaaraoui, P. Pérez and F. F. Revuelta, “A review on vision techniques applied to Human Behaviour Analysis for Ambient-Assisted Living,” *Expert Systems with Applications,* vol. 39, pp. 10873-10888, 2012.

[20] S. Vishwakarma and A. Agrawal, “A survey on activity recognition and behavior understanding in video surveillance,” *The Visual Computer,* vol. 29, no. 10, pp. 983-1009, 2012.

[21] M. Ziaeefard and R. Bergevin, “Semantic human activity recognition: A literature review,” *Pattern Recognition,* vol. 48, no. 8, pp. 2329-2345, 2015.

[22] D. Weinland, E. Boyer and R. Ronfard, “Action Recognition from Arbitrary Views using 3D Exemplars,” in *IEEE 11th International Conference on Computer Vision (ICCV),* Rio de Janeiro, 2007.

[23] H. C. Lim, E. Vats and C. S. Chan, “Fuzzy human motion analysis: A Review,” *Pattern Recognition,* vol. 48, no. 5, pp. 1773-1796, 2015.

[24] P. Afsar, P. Cortez and H. Santos, “Automatic visual detection of human behavior: A review from 2000 to 2014,” *Expert Systems with Applications*, vol. 42, no. 20, pp. 6935-6956, 2015.

[25] B. Zhang, N. Conci and F. G. B. D. Natale, “Camera viewpoint change detection for interaction analysis in TV shows,” *IEEE International Conference on Image Processing (ICIP),* pp. 2547 - 2551, 2014.

[26] K. Singh, R. Kapoor and K. S. Sinha, “Enhancement of low exposure images via recursive histogram equalization algorithms,” *Optik - International Journal for Light and Electron Optics*, vol. 126, no. 20, pp. 2619-2625, 2015.

[27] D. Weinland, M. Özuysal and P. Fua, “Making Action Recognition Robust to Occlusions and Viewpoint Changes,” in *11th European Conference on Computer Vision (ECCV) Part-III*, Greece, 2010.

[28] M. A. R. Ahad, J. K. Tan, H. Kim and S. Ishikawa, “Motion history image: its variants and applications,” *Machine Vision and Applications*, vol. 23, no. 2, pp. 255-281, 2012.

[29] D. K. Vishwakarma and R. Kapoor, “Hybrid classifier based human activity recognition using the silhouette and cells,” *Expert Systems with Applications*, vol. 42, no. 20, pp. 6957-6965, 2015.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] M. Piccardi , “Background subtraction techniques: a review,” in *IEEE International Conference on Systems, Man and Cybernetics*, Netherlands, 2004.

[35] L. Maddalena and A. Petrosino, “A Self-Organizing Approach to Background Subtraction for Visual Surveillance Applications,” *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1168-1177, 2008.

[36] S. Brutzer, B. Hoferlin and G. Heidemann, “Evaluation of Background Subtraction Techniques for Video Surveillance,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Providence, RI, 2011.

[37] T. S. F. Haines and T. Xiang, “Background Subtraction with DirichletProcess Mixture Models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 36, no. 4, pp. 670 - 683, 2014.

[38] A. Sobral and A. Vacavant, “A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos,” *Computer Vision and Image Understanding,* vol. 122, pp. 4-21, 2014.

[39] A. . H. S. Lai and N. H. C. Yung, “A fast and accurate scoreboard algorithm for estimating stationary backgrounds in an image sequence,” in *IEEE International Symposium on Circuits and Systems (ISCAS),* Monterey, CA, 1998.

[40] T. W. Chua, Y. Wang and K. Leman, “Adaptive Texture-Color based Background Subtraction for Video Surveillance,” in *19 th IEEE International conference on image processing (ICIP),* Orlando, FL, 2012.

[41] H. Zhang and D. Xu, “Fusing Color and Texture Features for Background Model,” in *Third International Conference on Fuzzy Systems and Knowledge Discovery,* Xi’an, China, 2006.

[42] C. H. Yeh, C. Y. Lin, K. Muchtar and W. L. Kang, “Real-time background modeling based on a multi-level texture description,” *Information Sciences,* vol. 269, no. 10, pp. 106-127, 2014.

[43] P. Chiranjeevi and S. Sengupta, “Neighborhood Supported Model Level Fuzzy Aggregation for Moving Object Segmentation,” *IEEE Transactions on Image Processing,* vol. 23, no. 2, pp. 645-657, 2014.

[44] J. C. Martínez, P. M. Martínez-Jim, J. M. Soto-Hidalgo and A. L. Salas, “A fuzzy approach for modelling visual texture properties,” *Information Sciences,* vol. 313, pp. 1-21, 2015.

[45] C. R. Wren, A. Azarbayejani, T. Darrell and A. P. Pentland, “Pfinder: Real-Time Tracking of the Human Body,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 780 - 785, 1997.

[46] C. Stauffer and W. E. L. Grimson , “Adaptive background mixture models for real-time tracking,” in *IEEE Conference on Computer Vision and Pattern Recognition,* Fort Collins, CO, 1999.

[47] A. Elgammal, D. Harwood and L. Davis, “Non-parametric Model for Background Subtraction,” in *6th European Conference on Computer Vision-Part II (ECCV),* London, UK, 2000.

[48] C. Lei and Y. H. Yang, “Optical Flow Estimation on Coarse-to-Fine Region-Trees using Discrete Optimization,” in *IEEE International Conference on Computer Vision,* Kyoto, 2009.

[49] H. Jiang, G. Zhang, H. Wang and H. Bao, “Spatio-Temporal Video Segmentation of Static Scenes and Its Applications,” *IEEE Transactions on Multimedia,* vol. 17, no. 1, pp. 3-15, 2015.

[50] J. L. Barron, D. J. Fleet and S. S. Beauchemin, “Performance of optical flow techniques,” *International Journal of Computer Vision,* vol. 12, no. 1, pp. 43-77, 1994.

[51] D. Fortun, P. Bouthemy and C. Kervrann, “Optical flow modeling and computation: A survey,” *Computer Vision and Image Understanding,* vol. 134, pp. 1-21, 2015.

[52] D. Sun, S. Roth and M. J. Black, “A Quantitative Analysis of Current Practices in Optical Flow Estimation and the Principles Behind Them,” *International Journal of Computer Vision,* vol. 106, no. 2, pp. 115-137, 2014.

[53] B. D. Lucas and T. Kanade, “An Iterative Image Registration Technique with an Application to Stereo Vision,” in *DARPA Image Understanding Workshop*, 1981.

[54] B. K. P. Horn and B. G. Schunck , “Determining optical flow,” *Artificial Intelligence,* vol. 17, no. 1-3, pp. 185-203, 1981.

[55] D. Weinland(a), R. Ronfard and E. Boyer, “A survey of vision-based methods for action representation, segmentation and recognition,” *Computer Vision and Image Understanding,* vol. 115, no. 2, pp. 224-241, 2011.

[56] M. M. Rahman and S. Ishikawa, “Robust appearance-based human action recognition,” in *IEEE International Conference on Pattern Recognition*, 2004.

[57] X. Dong, A.-C. Tsoi and S.-L. Lo, “Superpixel appearance and motion descriptors for action recognition,” in *International Joint Conference on Neural Networks (IJCNN),* Beijing, 2014.

[58] R. Melfi, S. Kondra and A. Petrosino, “Human activity modeling by spatio temporal textural appearance,” *Pattern Recognition Letters,* vol. 34, no. 15, pp. 1990-1994, 2013.

[59] E. Shechtman and M. Irani, “Space-Time Behavior-Based Correlation-OR-How to Tell If Two Underlying Motion Fields Are Similar Without Computing Them?,” *IEEE Transactions onPattern Analysis and Machine Intelligence,* vol. 29, no. 11, pp. 2045-2056, 2007.

[60] P. Dollar, V. Rabaud, G. Cottrell and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” in *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, 2005.

[61] B. Chakraborty(a), M. B. Holte, T. B. Moeslund, J. Gonzalez and F. X. Roca, “A Selective Spatio-Temporal Interest Point Detector for Human Action

Recognition in Complex Scenes,” in *IEEE International Conference on Computer Vision (ICCV)*, 2011.

[62] B. Chakraborty, M. B. Holte, T. B. Moeslund and J. Gonzàlez, “Selective spatio-temporal interest points,” *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 396-410, 2012.

[63] I. Everts, J. V. Gemert and T. Gevers, “Evaluation of Color Spatio-Temporal Interest Points for Human Action Recognition,” *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1569-1580, 2014.

[64] I. Jargalsaikhan, S. Little, C. Direkoglu and N. E. O’Connor, “Action recognition based on sparse motion trajectories,” in *IEEE International Conference on Image Processing*, Melbourne, VIC, 2013.

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

[66] M. S. Ryoo and J. K. Aggarwal, “Spatio-Temporal Relationship Match:Video Structure Comparison for Recognition of Complex Human Activities,” in *IEEE 12th International Conference on Computer Vision (ICCV)*, Kyoto, 2009.

[67] J. Hu and N. V. Boulgouris, “Fast human activity recognition based on structure and motion,” *Pattern Recognition Letters*, vol. 32, no. 14, pp. 1814-1821, 2011.

[68] B. Yin, W. Qi, Z. Wei and J. Nie, “Indirect human activity recognition based on optical flow method,” in *5th IEEE International Congress on Image and Signal Processing (CISP),* Sichuan, China, 2012.

[69] S. Ali and M. Shah, “Human Action Recognition in Videos Using Kinematic Features and Multiple Instance Learning,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 32, no. 2, pp. 288 - 303, 2010.

[70] S. Poularakis, K. Avgerinakis, A. Briassouli and I. Kompatsiaris, “Computationally efficient recognition of activities of daily living,” in *IEEE International Conference on Image Processing (ICIP),* Quebec City, QC, 2015.

[71] A. F. Bobick and J. W. Davis, “The recognition of human movement using temporal templates,” *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 23, no. 3, pp. 257-267, 2001.

[72] A. Chaaraoui, P. C. Perez and F. Revuelta, “Sihouette-based human action recognition using sequences of key poses,” *Pattern Recognition Letters,* vol. 34, pp. 1799-1807, 2013.

[73] A. A. Chaaraoui and F. F. Revuelta, “Optimizing human action recognition based on a cooperative coevolutionary algorithm,” *Engineering Applications of Artificial Intelligence,* vol. 31, pp. 116-125, 2014.

[74] 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.

[75] D. N. Olivieri, I. G. Conde and V. X. A. Sobrino, “Eigenspace-based fall detection and activity recognition from motion templates and machine learning,” *Expert Systems with Applications,* vol. 39, no. 5, pp. 5935-5945, 2012.

[76] A. Eweiwi, S. Cheema, C. Thurau and C. Bauckhage, “Temporal key poses for human action recognition,” in *IEEE International Conference on Computer Vision Workshops (ICCV Workshops),* 2011.

[77] 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.

[78] C. Thurau and V. Hlavac, “Pose primitive based human action recognition in videos or still images,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[79] J. Dou and J. Li, “Robust human action recognition based on spatio-temporal descriptors and motion temopral templates,” *Optik - International Journal for Light and Electron Optics,* vol. 125, no. 7, pp. 1891-1896, 2014.

[80] L. Shao, R. Gao, Y. Liu and H. Zhang, “Transform based spatio-temporal descriptors for human action recognition,” *Neurocomputing*, vol. 74, no. 6, pp. 962-973, 2011.

[81] I. T. Jolliffe, Principal Component Analysis, Springer-Verlag New York, 2002.

[82] O. Masoud and N. Papanikolopoulos, “A method for human action recognition,” *Image and Vision Computing*, vol. 21, no. 8, pp. 729-743, 2003.

[83] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of Human Genetics*, vol. 7, no. 2, pp. 179-188, 1936.

[84] T. Cover and P. Hart, “Nearest neighbor pattern classification,” *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21-27, 1967.

[85] 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.

[86] S. Sadek, A. A. Hamadi, M. Elmezain, B. Michaelis and U. Sayed, “Human Action Recognition via Affine Moment Invariants,” in *21st International conference on Pattern Recognition*, 2012.

[87] C. W. Hsu and C. J. Lin, “A Comparison of Methods for Multiclass Support Vector Machines,” *IEEE Transactions of Neural Network*, vol. 13, no. 2, pp. 415-425, 2002.

[88] H. Qian, Y. Mao, W. Xiang and Z. Wang, “Recognition of human Activities Using SVM Multi-Class Classifier,” *Pattern Recognition*, vol. 31, no. 2, pp. 100-111, 2010.

[89] C. C. Chang and C. J. Lin, “A library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology,* vol. 2, no. 3, pp. 1-27, 2011.

[90] 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, 2013.

[91] M. A. Tahir, A. Bouridane and F. Kurugollu, “An FPGA Based Coprocessor for GLCM and Haralick Texture Features and their Application in Prostate Cancer Classification,” *Analog Integrated Circuits and Signal Processing*, vol. 43, no. 2, pp. 205-215, 2005.

[92] R. Haralick, K. Shanmugam and I. Dinstein, “textural Features for Image Classification,” *IEEE trans. on Systems Man, and Cybernetics*, vol. 6, pp. 610-621, 1973.

[93] G. Guo and A. Lai, “A survey on still image based human action recognition,” *Pattern Recognition,* vol. 47, no. 10, pp. 3343-3361, 2014.

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

[95] D. K. Vishwakarma, R. Kapoor and A. Dhiman, “A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics,” *Robotics and Autonomous Systems*, vol. 77, pp. 25-38, 2015.

[96] V. N. Vapnik, “An overview of statistical learning theory,” *IEEE Transactions on Neural Network,* vol. 10, no. 5, pp. 989-999., 1999.

[97] C. Schuldt, I. Laptev and B. Caputo, “Recognizing human actions: a local SVM approach,” in *Proc. of the International conference on Pattern Recognition*, 2004.

[98] A. Fathi and G. Mori, “Action Recognition by Learning Mid-level Motion Features,” in *Proc. of IEEE conference on Computer Vision and Pattern Recognition,* 2008.

[99] R. Touati and M. Mignotte, “MDS-Based Multi-Axial Dimensionality Reduction Model for Human Action Recognition,” in *Proc. of IEEE canadian Conference on Compter and Robot Vision*, 2014.

[100] 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.

[101] S. Rahman, I. Song, M. K. H. Leung and I. Lee, “Fast action recognition using negative space features,” *Expert Systems with Applications*, vol. 41, pp. 574-587, 2014.

[102] I. G. Conde and D. N. Olivieri , “A KPCA spatio-temporal differential geometric trajectory cloud classifier for recognizing human actions in a CBVR system,” *Expert Systems with Applications*, vol. 42, no. 13, pp. 5472-5490, 2015.

[103] Y. Wang and G. Mori, “Human Action Recognition by Semi-Latent Topic Models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1762-1764, 2009.

[104] 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.

[105] X. L. Ming, H. J. Xia and T. L. Zheng, “Human action recognition based on chaotic invariants,” *Journal of South Central University*, vol. 20, pp. 3171-3179, 2013.

[106] 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.

[107] L. Li and M. K. H. Leung, “Integrating intensity and texture differences for robust change detection,” *IEEE Transactions on Image Processing*, vol. 11, no. 2, pp. 105-112, 2002.

[108] J. Han and B. Bhanu, “Individual recognition using gait energy image,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 28, no. 2, pp. 316-322, 2006.

[109] A. Bosch, A. Zisserman and X. Munoz, “Representing shape with a spatial pyramid kernel,” in *ACM International conference on Image and Video Retrieval*, Amsterdam, 2012.

[110] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *IEEE Conference onComputer Vision and Pattern Recognition (CPVR)*, San Diego, CA, USA, 2005.

[111] H. Zhang, Z. Liu and H. Zhao, “Recognizing human activities by key frame in video sequence,” *Jounral of Software,* vol. 5, no. 8, pp. 818-825, 2010.

[112] Z. Khan and W. Sohn, “Abnormal human activity recogntion system based on R-Transform and Kernel Discriminant Technique for Eldely Home Care,” *IEEE Transactions on Consumer Electronics,* vol. 57, no. 4, pp. 1843-1850, 2011.

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

[114] M. Singh, M. Mandal and A. Basu, “Pose Recognition using the Radon Transform,” in *48th Midwest symposium on Circuits and Systems*, 2005.

[115] Y. Wang, K. Huang and T. Tan, “Human Activity Recognition Based on R-Transform,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

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

[117] 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.

[118] J. Nibeles, H. Wang and L. Fei-Fei, “Unsupervised learning of human action categories using spatial-temporal words,” *international Journal of Computer Vision,* vol. 79, no. 3, pp. 299-318, 2008.

[119] A. Kläser, M. Marszałek and C. Schmid, “A spatio-temporal descriptor based on 3d-gradients,” in *Proceedings of the British Machine Vision Conference (BMVC’08),* Leeds, United Kingdom, 2008.

[120] T. Y. Hung, . L. Jiwen, H. Junlin, Y. P. Tan and G. Yongxin, “Activity-based human identification,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Vancouver, BC, 2013.

[121] R. Benmokhtar, “Robust human action recognition scheme based on high-level feature fusion,” *Multimedia Tools and Applications,* vol. 69, no. 2, pp. 253-275, 2014.

[122] B. Yao, H. Hagras, M. J. Alhaddad and D. Alghazzawi, “A fuzzy logic-based system for the automation of human behavior recognition using machine vision in intelligent environments,” *Soft Computing,* vol. 19, no. 2, pp. 499-506, 2015.

[123] D. K. Vishwakarma and R. Kapoor, “Integrated Approach for Human Action Recognition using Edge Spatial Distribution, Direction Pixel, and R - Transform,” *Advanced Robotics,* vol. 29, no. 23, pp. 1551-1561, 2015.

[124] 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.

[125] C. N. Doukas and I. Maglogiannis, “Emergency Fall Incidents Detection in Assisted Living Environments Utilizing Motion, Sound, and Visual Perceptual Components,” *IEEE Transactions on Information Technology in Biomedicine,* vol. 15, no. 2, pp. 277-289, 2011.

[126] Â. Costa, J. C. Castillo, P. Novais, A. Fernández-Caballero and R. Simoes, “Sensor-driven agenda for intelligent home care of the elderly,” *Expert Systems with Applications,* vol. 39, no. 15, pp. 12192-12204, 2012.

[127] M. Bosch-Jorge, A.-J. Sánchez-Salmerón, Á. Valera and C. Ricolfe-Viala, “Fall detection based on the gravity vector using a wide-angle camera,” *Expert Systems with Applications,* vol. 41, no. 17, pp. 7980-7986, 2014.

[128] 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.

[129] Cheng Yang, U. Ugbolue, B. Carse, V. Stankovic, L. Stankovic and P. Rowe, “Multiple marker tracking in a single-camera system for gait analysis,” in *IEEE International Conference on Image Processing (ICIP)*, Melbourne, VIC, 2013.

[130] A. Procházka, M. Schätz, O. Ťupa, M. Yadollahi, O. Vyšata and M. Walls, “The MS kinect image and depth sensors use for gait features detection,” in *IEEE International Conference on Image Processing (ICIP)*, Paris, 2014.

AUTHOR BIOGRAPHY

	Dinesh Kumar Vishwakarma, Assistant Professor, Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India Email: dvishwakarma@gmail.com, dkvishwakarma@dce.ac.in

Dinesh Kumar Vishwakarma received the Bachelor of Technology (B.Tech.) from Dr RML Avadh University, Faizabad, Uttar Pradesh India, in the year 2002 and the Master of Technology (M.Tech.) from Motilal Nehru National Institute of Technology, Allahabad, Uttar Pradesh, India in the year 2005. Presently, he is working as an Assistant Professor, in the Department of Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India-110042. His research interests include computer vision, pattern recognition, human action and activity recognition, and hand gesture recognition.