A THESIS REPORT
ON

"HUMAN ACTION RECOGNITION BASED ON R-TRANSFORM AND SPATIAL TEMPORAL DESCRIPTOR"

Submitted in partial fulfillment of the requirement for the award of the degree of

Master of Technology
In
Signal Processing & Digital Design

Submitted by

Ashish Dhiman

2K12/SPD/03
Under the Supervision of

Mr. Dinesh Kr. Vishwakarma

(Assistant Professor)

DEPARTMENT OF ELECTRONICS & COMMUNICATION ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
Shahabad Daulatpur, Bawana Road, Delhi-110042, India
JULY, 2014

Certificate

Date: ___/___/_____

This is to certify that the thesis ***"Human Action Recognition based on R-Transform and Spatial Temporal Descriptor"*** is the authentic work of ***Mr. Ashish Dhiman*** under my guidance and supervision in the partial fulfillment for the Degree of ***Master of Technology*** in ***Signal Processing & Digital Design*** to Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India. To the best of my knowledge, the matter embodied in the thesis has not been submitted to any other University/ Institute for the award of any other degree.

(D.K. Vishwakarma)
Supervisor, Assistant Professor,
Department of Electronics and Communication Engineering,
Delhi Technological University, Delhi-110042

Acknowledgement

With all praises to the almighty and by His blessings I have finally completed this thesis.

I would like to express my gratitude to ***Sh. Dinesh Kr. Vishwakarma****, Assistant Professor, Department of Electronics & Communication Engineering, Delhi Technological University, Delhi, who has graciously provided me his valuable time whenever I required his assistance. His counseling, supervision and suggestions were always encouraging and it motivated me to complete the job in hand. He will always be regarded as a great mentor for me.*

I also offer my sincere thanks to ***Prof. Rajiv Kapoor****, Head of Department (Electronics and Communication Engineering), Delhi Technological University for introducing me to the world of computer vision. I learnt from him not only the foundation knowledge and state-of-the-art research, but also his dedication and persistence toward research.*

I have no words to express my gratitude to my parents, ***Sh. P.C. Dhiman*** *and* ***Smt. Neelam Kumari*** *and brother* ***Mr. Aditya Dhiman*** *for nurturing me with love, with the curiosity for learning and research, for being my inspiration in every dimension of life, and for giving me the strength to carry on during the hard times of this journey.*

Finally, I would like to thank my friends for their love, support and encouragement that reinforced my spirits at some crucial junctures.

Ashish Dhiman
Roll No: 2K12/SPD/03
M. Tech.
(Signal Processing & Digital Design)
Department of Electronics &Communication Engineering,
Delhi Technological University, Delhi - 110042

Abstract

This thesis presents the framework of the combination of motion and shape information produced by different actions for human action recognition. The motion information obtained from the Radon Transform, computed on the Binary Silhouettes obtained from video sequence. Radon Transform of an image gives the projection of lines in all directions therefore it gives information about pixels variation inside the shape. We use the properties of Radon transform i.e. invariance to scaling and translational, to make noise free and robust model and rotational variance to distinguish the different actions. For the shape information we generate the set of static shapes such as MHI/MEI and AEI. Features are extracted from these models by using Pyramid of Histograms (PHOG), Directional Pixels and 2-D DFT. PHOG represents the concatenation of Histogram of Gradients (Hog) over the sub-regions and will give the global spatial information about the shape representation. Shape feature vectors alone will not give discriminating features to recognize the actions such as "run" and "walk", "jumping at one place" and jumping forward, therefore we integrate with motion descriptor that provides the angular variations while performing actions. We also present another integrated model with motion descriptor where we use Single action images for the representation of pose of action. Single action images are extracted from the sequence of videos using the Fuzzy inference system. Finally with all the extracted features, we train the system using a Support Vector Machine and K-Nearest Neighbour algorithm to recognize the various actions. Weizmann dataset is used for Evaluation.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

2D.....Two Dimensional
3D…….……..Three-dimensional
HAR.....Human action recognition
MHI…….Motion History
MEI……..…Motion Energy
AEI…….……Action Energy Image
PHOG……...…Pyramid of histogram
ROI.....Region of Interest
DFT……...…Discrete Fourier transform
KNN……...……..K-nearest neighbour
SVM……...…Support vector machine
SIFT…….……Scale invariant feature transform
HOF…….……...…Histogram of Optical flow
PLSA……Probabilistic Latent semantic analysis
LDA……..…….….Latent Dirichlet allocation
HOG.....…...Histogram of Oriented Gradient
STSP……...….Spatial-temporal steerable pyramid
COP.....…...Cloud of Interest points
BOW……..……..…Bag of words
VMT…….……Volume motion template
PMT…….....…Projected motion templates
HTT…….....…History Trace Templates
HTF…….……History Triple Features
PMF…….…..……. Pyramidal Motion feature
GEI…….……..…Gait energy image
HOG…….……..…Histogram of Oriented Gradient
RT……...…….……Radon Transform
SMM……...……...……..…Spatial Maxima mapping
LDA……....Linear Discriminant analysis
DR.....…...Dimension Reduction
PCA…….Principal Component Analysis
LLE…….…….……Local Linear Embedding
HOI.....…...Human object Interaction

Chapter1

1.1 Introduction

Studying and analyzing human action from the videos is interesting and appealing task, yet among the difficult problems of computer vision. The answer to this problem will serve many important applications ranging from surveillance to interaction between the computer and humans. Surveillance of the public using network cameras is common in many areas around the world. Burgeoning of IP cameras used for visualizing human actions has led to surveillance video overload. But computation complexities and the volume of information produced by them exceeds the capacity of the human analyst to analyze, debug and respond in real time. Therefore, Authentic, functions as well as efficacious automatic algorithms using signal processing techniques are required to analyze the visual environment. Human action recognition (HAR) is nothing but recognition of human actions based on videos. Actions can be explained as single human behaviour analysis, such as running, walking, punching. The basic human action recognition model consists of three main processing steps i.e., Object segmentation, feature extraction from the object characteristics and then recognize these actions.

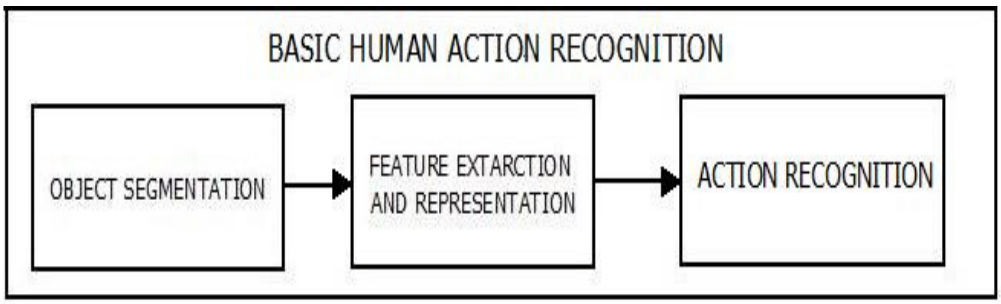


Fig 1.1: Basic Human Action Recognition System

There are various algorithms which already for the object segmentation like background subtraction, Gaussian mixture model, temporal difference, optical flow.

Segmentation of the object depends upon the camera movement, whether it is stationary or moving. When the camera is static then we can easily obtain the object from the video sequence by applying background subtraction. In case of background subtraction we first find the background image and then in other frames we locate the object. From the video sequence, we subtract the background image frame by frame

from the object. Gaussian mixture model is most useful when there are more illumination changes in the video sequence. It is based on the concept of pixel probability, higher the probability implies that it belongs to the background. Optical flow and temporal method are applied when our background is not varying. In case of temporal differencing, we subtract the consecutive frames (T and T+1) and finds the object. Optical flow finds the displacement changes of pixels corresponding to the object. The complexity of this method increases in case of occluded images. This chapter contains a general introduction to the problem and motivation for this thesis. Finally, we describe our proposed method and outline of the thesis.

1.2 Motivations

The interest in human activity recognition topic is inspired by many promising applications which we use in day to day life.

1) **Video Surveillance:** Keeping public safety and privacy protection into consideration, large number of cameras and monitors are deployed in public areas such as city malls, airports, railway stations and bus stands. They produce endless video streams. Out of them, one with unusual actions is used by human operators and other investigating authorities. So to avoid extra human effort we have algorithms which distinguish videos based on unusual actions and only distinguished one can be used for analysis.
2) **Monitoring crowded scenes:** Consider an example, if there is fire at city mall or if we discover a flock or we can say mob running away from a place or a building or if we monitor people fighting etc. Then all these scenarios indicate that something wrong happened there. Therefore, with the help of automated system an alarm can be automatically raised to the emergency services for the help.
3) **Sports video analysis and evaluation:** In sports we can learn the characteristics of the athletes like players body movements in the game or what techniques they follow. This kind of learning in sports helps a lot to athletes in visualizing and improving their skills.
4) **Gesture-Based Interactions:** Now a day's research is going on for improving the interaction between the human and physical environment. There are various interactive games are available in market without physical controller like any

interface medium such keyboard, joystick etc. They use voice and gestures and perform actions accordingly with the use of algorithms. The best example is the Microsoft Kinect Sensor.

5) **Biometrics Analysis:** Biometric means identifying the humans based on physical traits such as fingerprints or face recognition. These biometrics systems are installed in many societies, corporate offices, laptops etc. This system can process multiple queries and can identify multiple people without interrupting or interfering with the subjects activities. E.g. of behavioural biometrics is human gait recognition, face recognition, iris recognition.
6) **Intelligent Transport Systems for safety driving:** For Intelligent Transport Systems (ITS), action recognition is important and challenging task due to varying illumination and poor quality data processing images. Also, the presence of other moving vehicles as well as multiple people, mainly pedestrian makes it very difficult. Image depth analysis is also required. Large numbers of companies are spending funds on this. Moreover, involvement of moving cameras (within a vehicle) in a higher speed is another challenging area.
7) **Healthcare system:** Human action recognition can be used in monitoring the development of the child, observing the elderly person in the homes and plays an important role in hospitals where doctors continuously monitor the condition of the patient while sitting in some other place. Various algorithms are developed which study the human posture, movements, pattern of actions.
8) **Safety in driving:** For the safety of drivers while driving various algorithms are developed which study the driver's characteristics and help them against accidents. Automatic airbags, alarm system are installed in the modern vehicles, which help them in safe driving. For e.g. sometimes while driving, drivers feel lazy and feels distraction, in that case system will alert the driver through a voice processor alarm.
9) **3-D Animation:** In case of 3-D technology, human motion activities or images are recorded and then used for designing the 3-D animated games, movies and study the analysis of human action movements. This technology plays a great role in a HAR system as it provides more information about the characteristics of human than the 2-D image. E.g. the human skeleton model is used for estimating the movement of human motion.

1.3 Challenges

1) Optical flow, point trajectories, space-time gradients, and sparse interest points are common features used in action recognition analysis. In case, if the video quality is low and motion is not smooth then optical flow vectors can result in inaccurate analysis. Similarly, in case of point trajectories, one needs proper tracking method. In case of interest point representation, the distribution of points should be stable. This representation does not provide the scale and translation. Silhouette representation of actions provides informative features for recognition and invariant to illumination changes. But the accuracy of recognition depends upon the background subtraction techniques which may face problems in multiple human activities, occlusions and dynamic background.
2) Other than these, with the use of large number of cameras in public places such as airports, railway stations, city malls etc. Data volume is increasing exponentially. Lots of video sequences with different angles are available, so this makes it very difficult to synchronize and recognize for human operators. Videos having cluttered and occlusions are less efficient for action recognition. Moreover, lighting conditions vary to a large extent and these variations in lighting make the recognition process a tougher one.
3) If we evaluate the condition, practically, intra-class motion variability exists i.e. Different individuals perform same action differently. Similarly, different actions may look similar in some particular poses. It is known as inter-class motion ambiguity such as jumping and skipping.
4) Different datasets are used by different methods and as such no particular algorithms are there to determine which dataset is better than the other.
5) Some Data samples can be redundant in redundant in nature, whereas some actions can also have an associated periodicity of movements as actions mostly vary in terms of dimensionality and overlapping. For example, waving hands, running, continuous standing and sitting actions, etc. So it becomes very difficult to recognize for redundancy as well as overlapping patches.
6) For real time systems execution rate is another challenge. For a stream of videos Human operator cannot buffer among them as this is a tedious task and also time consuming. So we need algorithms that could search for unusual actions

and those particular videos should be further transmitted to a human operator for research or analyzing purpose.

7) Loose clothing, self-occlusions, inaccurate body models, unavailability of limbs, turns model-based approaches into jeopardy. So, recognition becomes difficult because one cannot model a moving actor in a reasonable manner. Therefore, plenty of avenues remain wide open or not filled yet.

1.4Proposed Methodology

This thesis will illustrate various models which will help to resolve the action recognition problem. The proposed framework based on the combination of appearance and motion features which are executed separately but sequentially. Human posture features give the information about the whole body movements. For the representation 2-Dimensional (2-D) shape we worked on the Motion History and Motion Energy (MHI/MEI) images.

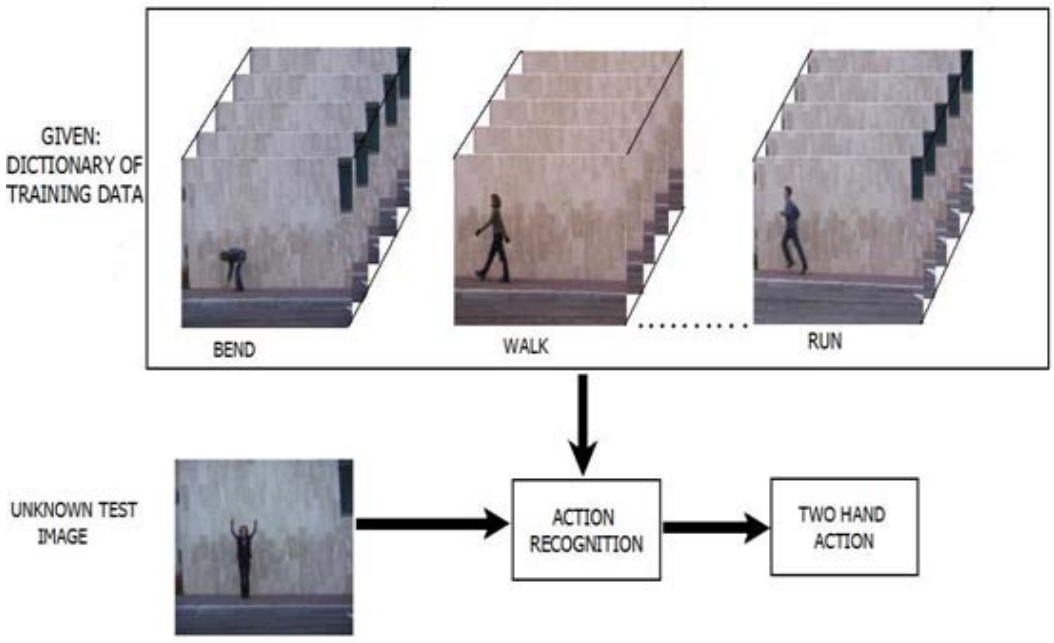


Fig 1.2: Overview of Action Recognition concept where we test the input image and system will give output about the action on the prior knowledge of training actions.

These images are easy to form and robust in action representation. MHI images will give the latest information about the motion of action and MEI will give not only type of motion but also the direction of motion. We further extend the concept of MHI/MEI images to Action Energy Image (AEI) which is more robust in representation. AEI

images remove the time dependency limitation and contain more information as compared to MHI/MEI images. AEI images will easily differentiate the similar representation actions of different categories on the basis of the intensity of pixel values. From the AEI images we try to solve the limitation of MHI/MEI images with improved results.

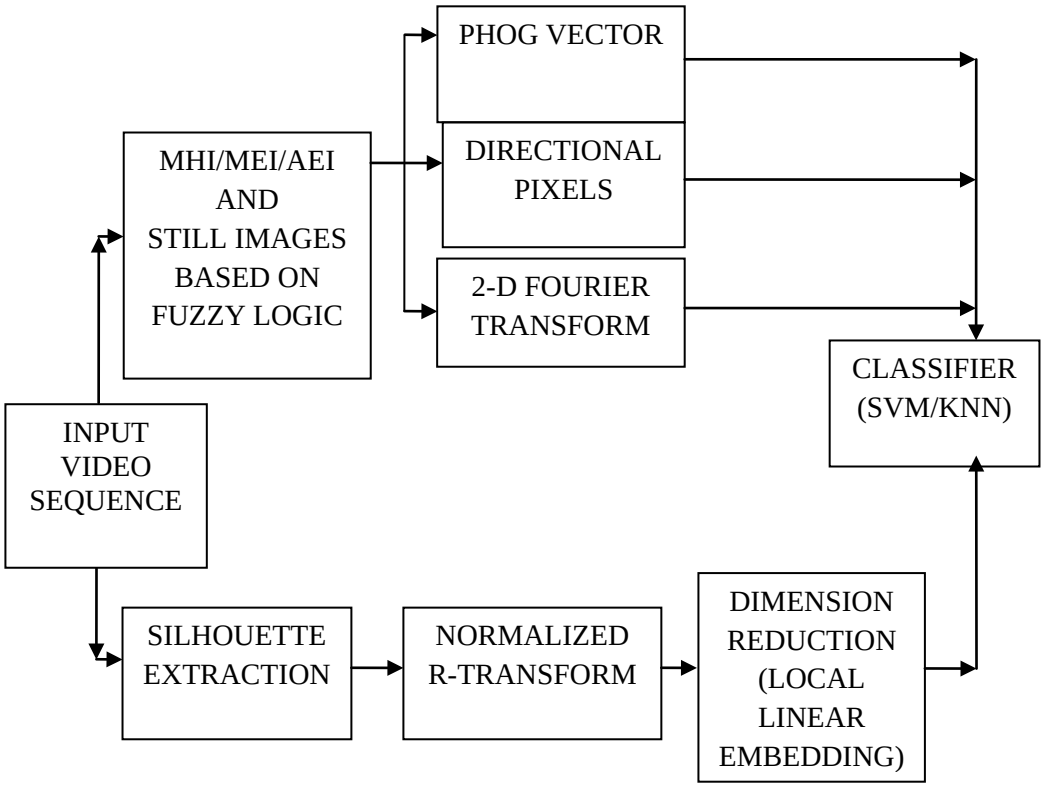


Fig 1.3: Overview of the proposed methodology

These 2-D representations of images give the global information about the human posture. We used Pyramid of histograms (PHOG) as a local descriptor to extract the features from the static images represented by MHI/MEI/AEI. PHOG is based on the concept of the histogram of gradients, but in case of PHOG we divide the image into sublevels and for each sub region we compute the orientation of the edges. In PHOG, we first segment out the Region of interest (ROI) from the static images. This helps us to remove the redundant information around the object i.e. Background and gives more structural information about the object. ROI is divided into further levels and then for each level, we find out the orientation of the edges. PHOG vector is the summation of all the oriented edges computed in each sub-levels. There is another two feature 1) Variation of Directional pixels 2) 2-D Fourier transform. Directional pixels will give more internal information of the static image and then we compute the Fourier

transform on these pixels, which will give changes frequency domain. These features will effectively differentiate the action of the same class with different class and captures both the spatial and temporal information. This representation is embedded with motion information obtained from the Binary Silhouettes.

We used the Radon transform for the computation of motion features. This is further extended to Normalize R-Transform which gives invariance to translation and scaling and finds the angular variation in the actions. It will give another discriminative feature for recognizing the human actions. This proposed system has sufficient discriminative combined descriptors that will give us robust and improved accuracy.

We also proposed the idea of combination of motion features with still images. Still images are extracted from the video by using Fuzzy inference system. These images do not require any background subtraction or morphological operation; therefore it reduces the computation time and complexity of the system. In this thesis, we worked on the single human action video dataset i.e. Weizmann dataset having activities like walking, bending, jumping, waving, running, all of which can intuitively be described by features that represent the human as a single object.

1.5 Outline of the Thesis

The outline of the thesis is as follows:-

Chapter 2 reviews literature survey on action recognition. It categorizes past work with respect to various action representation models, feature extraction techniques and their advantages and disadvantages.

Chapter 3 introduces our first step of proposed action recognition methodology i.e.2D posture representation of actions performed in the video sequence. First, we describe the formation of MHI/MEI and AEI images which give a static posture representation of an object. After that we explain our feature extraction techniques that we performed on our static posture images. We used three feature extraction techniques named PHOG, Directional pixel variations along the x-y direction and 2-DFT (Discrete Fourier transforms) used to identify the actions.

Chapter 4 discusses the representation of our Motion based descriptor, R- Transform that is used to analyze the action motion in video sequences. As the dimensions of

Radon feature vectors are large we used the Dimension Reduction technique which gives us the compact representations of features vectors.

Chapter 5 discusses the brief introduction about Dimension Reduction technique and their mathematical algorithm.

Chapter 6 in this chapter the concept of Still images is explained. For the extraction of still images, we formulate the Key frame extraction algorithm which is based on the Fuzzy logic inference system. The concept of Fuzzy logic and algorithm related to Key frames are also explained.

Chapter 7 discusses the action classification algorithms. We used the Multi-class Support vector machine (SVM) and K-nearest neighbours (KNN) classifier with leave one out approach for action recognition.

Chapter 8 discusses the experiment and the results of our proposed method. This chapter also discusses the limitation of methods.

Chapter 9 At last, we describe the conclusions and presents possible ways for future work.

Chapter2

Literature Review

This section briefly describes the previous work in Human Action Recognition (HAR) system and their feature extraction models with advantages and limitations.

Basic Feature Extraction Models of HAR

In general, we divide the HAR feature extraction models into global, local and body (2D-3D) approaches. Global approaches work on silhouette images or contour of the human body which gives the shape information, whereas local representation works on the small patches which is used in object recognition. With the invention of Microsoft Kinect, 3D modelling of human actions is easy and in better resolution compared to the 2D representation.

2.1 Body Approaches

Body approaches describe the 2D and 3D representation of the human posture. The 2D representation of human motion works on the overall information about action image, whereas 3D will have another feature information i.e. Depth, which means the distance of the object from the camera. This 3D representation is helpful in overcoming the limitation of occlusion and gives more accuracy compared to 2D representation and also differentiates the similar actions like 'eating food' and 'drinking a cup of tea'. Li et al. [1] presented the set of 3D points on the depth silhouettes which are represented as the feature descriptors. They used the Hausdorff distance to find out the dissimilarity between the depth maps. This method gives lower recognition accuracy due to the presence of noise and occlusion in depth maps. Raptis et al. [2] used the concept of joint angles in skeleton tracking to recognize the various dance actions. This method of representation is limited to a small number of classes. Yang et al. [3] introduced the new feature vector called as "Eigen Joints" based on depth maps which efficiently represent the both posture and motion features. They transform the depth maps into Depth motion maps by dividing into 3 axis i.e. *x-y*, *y-z*, *z-x* and used the HOG to recognize the action. In the end, Naive Bayes classifier is used for the multi-class action classification. Wang et al. [4] introduced the concept of 'actionlet ensemble model'. They form the clusters of joints and the depth neighbourhood of similar action in one group. This method shows the advantage in case of intra-class variations. Tran et

al. [5] used the polar histogram approach for the representation of skeletonization.Vieira et al. [6] proposed the new feature vector ‘Space–Time Occupancy Patterns’ for classifying the human action from depth maps. Depth maps are divided intogrids4D grid is obtained by dividing the depth image into time axis. PCA is employed to reduce the dimension of STOP feature vector because of its sparse representation. This method is more flexible and shows good accuracy compared to others. J. Luo et al. [7] proposed the approach based on RGB-D images. They introduced Center-Symmetric Motion Local Ternary Pattern (CS-MLTP) approach to extract both the spatial and temporal features from the RGB sequences. They also introduce sparse coding on extracted features which improves the accuracy of the system. This method outperforms in actions having large variations. Wang et al. [8] proposed the novel actionlet ensemble model for action representation based on the correlation of human joints. They introduce the new temporal representation based on a Fourier temporal pyramid and the local occupancy pattern in depth information. The proposed model is invariant to translation, temporal misalignment and robust to noise.

2.2 Local Approaches

Lowe et al. [9] proposed the concept of Scale invariant feature transform (SIFT) which varies with illumination changes, but invariant to scale, rotation and translation changes. Sift has some drawbacks 1) high dimension vector, 2) misses the colour information (RGB), 3) it will not discriminate the background and foreground easily. Histogram of oriented gradients is most popular approach for the object detection. Dalal and Triggs [10] proposed the HOG for the object detection in still images. Felzenszwalb et al. [11] presented method for human detection based on part based models. They modelled the human into various sub parts and for each subpart HOG are computed. This is better and fast representation of a human detection. Mathieu et al. [12] used the histogram to recognize the ongoing human activity. They extend the use of classical HOG into small sub-parts histograms. Each action is sliced into sub-parts histograms which make it more flexible and efficient representation of the histogram in recognizing action. The main limitation of HOG features is that human size should not vary, its scale remains fixed. Onofri et al. [13] used the combination HOG with Histogram of optical flow (HOF) for the classification of human action based on the subsequences. The video is divided into a number of small sub sequences. For the

feature extraction they introduced the MOSIFT descriptor which is the combination of HOG and Histogram of Optical flow (HOF) applied to each subsequent. Then visual vocabulary is formed of all STIP's descriptors based on K-means algorithm for the size reduction SVM classifier is used for classification and it provides better accuracy while working on subsequence. This method is stable and robust in representation compared to part based models.

Laptev et al. [14] introduced the concept of interest points. They analyze the image in x-y-t dimension and detect the presence of interest points in space time volumes. Their method of generation of interest points is not stable and effective in complex actions. Dollar et al. [15] worked on the limitations and proposed the interest points based on the separable linear filters. Gorelick et al [16] extended the 2D motion template to 3D space time volume. They extract features like space time saliency, action dynamics, shape structure and orientation based on the Poisson equation solution. Space time volume captures both spatial and temporal information. Laptev et al [17] improved his results by proposing his spatio-temporal histogram model. In this method histogram model segmented into small grids and locate the interest points in the grids which give the information about the motion in video sequences. Niebles et al. [18] presented the probabilistic Latent semantic analysis (PLSA) and Latent Dirichlet allocation (LDA) models over the space time regions to perform unsupervised action recognition. These models have the limitation as they do not provide temporal and scale invariance. Kovashka et al. [19] used the joint positions as a discriminating feature for action recognition. Trajectories of motion are formed from frame by frame variations of joint positions. The points are shown in the 3-D space, and it compares trajectory shapes to classify human actions. The joint positions on the graph can be used to determine the actions taken by a person. The advantage to this approach is the ability to track and analyze human movements. They are viewpoint dependent and require compact representation of interest points. Later on, Zhao et al. [20] presents the combined approach representing both structure and appearance information based on STIP's. For the appearance information they used cuboids which are extracted from the STIP's. They used local descriptor 3D Histogram of oriented gradients (HOG3D) as a feature vector for the appearance representation and for the structural information, Sphere kernel $s\,(r, \theta, \partial)$ are computed over STIP.Yi and Lin [21]used the trajectory saliency first time in human action recognition. These trajectories are obtained from the video

sequence depending upon the saliency of a region is computed by comparing the features with surrounding window regions. Kernel Histograms are computed over the trajectories having high saliency value. Zhen et al. [22] presents the spatial-temporal steerable pyramid (STSP) approach for human action recognition. They used a Laplacian pyramid approach to decompose the video sequence and then multilevel steerable filters are used to extract the features in different directions and scale. G. Somasundaram et al. [23] introduced spatio temporal feature detector based on sparse representation. They used the sparse representation to measure the saliency of the patches. On the salient patches they further computed the HOG and region covariance descriptors which form the code book based on Bag of features. This method works well for large data set and does not require large storage memory as it works on the saliency approach and sparse representation makes the system scale invariant. Shao et al. [24] introduced the novel Laplacian pyramid coding descriptor for the holistic representation of human action. For the representation of salient features whole video sequence is divided into series of band pass filtered components and then the Gabor filter is employed to extract the discriminative spatio-temporal features. This method is independent of tracking of features or localization of STIP's. The performance of trajectory based recognition methods dependent upon quality of trajectories.

However, these space time approaches is not suitable for recognizing multiple or more complex activities that are not periodic in nature, sensitive to partial occlusion, background variations. K.N. Tran et al. [25] proposed the part based model for recognition of human. First human parts are extracted from the subject in all the frames of input video and they are transformed into polar coordinate space that is quantized in both in distance and angular orientation. Polar representation used as the discriminating representation between the two different body parts motion. Classification is based on the sparse representation which improves the accuracy. This method is robust to partial occlusion and complex activities. M. Bregonzio et al. [26] represents the combination of appearance and distribution information of interest point for recognition of action. They represent the cloud of interest points (COP) accumulated at different scales. They extracted the two features COP and bag of words (BOW) from space-time interest's points. Multiple kernel learning (MKL) is used to find the discriminating features for the classification. Although this method shows improved results, but it is not applicable to dynamic background, multiple people performing activities.

2.3 Global approaches

Global approaches represent the appearance and motion of the actions based on the silhouettes and temporal models. The performance of Silhouette based models depends upon the background subtraction techniques. Therefore, it may face problems where the background is dynamic, occlusion and multiple people in action video. But Silhouettes are insensitive to illumination, colour variations. Template models represent the action taking whole video sequence rather than short duration of period.

Bobick and Davis [27] are the first to use the concept of template representation where they formed MHI/MEI templates for action recognition. They compute the 7 '*Hu*' moments on the images which they used as feature descriptors to describe the motion and in classification. The representation of MHI/MEI is simple, but they are dependent upon time parameter and view dependent. The formation of MHI is also dependent upon the variation of an action speed which degrades the performance of the recognition system. Efros et al [28] perform recognition on low resolution videos by correlating the optical flow measurements. M.C. Roh et al [29] proposed the volume motion template (VMT) and projected motion templates (PMT) view independent images. VMT represents the 3D view of MHI and PMT will help in projecting the VMT back in 2D plane. This method removes the view point dependency in case of [1]. L. Shao et al. [30] proposed the recognition of activities on the combination of motion and shape analysis. First they formed MHI images of the activities and then pyramid of correlogram (PCOG) applied for the shape description. PCOG represents the shape information through the gradients and the spatial distribution of gradients with neighbouring location. They also compared the PCOG with the descriptor like HOG AND HU moments and finds improved results. Goudelis et al. [31] presents the method based on Trace transform for action recognition. In this work they computed History Trace Templates (HTTs) and History Triple Features (HTFs) which give both spatio temporal information and invariant to different variations. Trace transform is computed on each binary silhouette image and template is formed by integrating all transform images. Compared to other method, it has complexity and also invariant to scaling, translation and rotation. C. P. Lee et al. [32] proposed the time sliced average MHI image with HOG in a gait recognition. Gait cycles are divided into several regular windows to generate the same number of Time sliced MHI. The hog is calculated for

the each Time sliced MHI which will give features for classification. This method is useful in the elimination of template variation.

Pose based recognition [33, 34, 35, 36] works on the still images for recognition. Therefore, they do not contain the motion information in short or long time duration. They do not involve the background subtraction techniques or tracking trajectories, robust to noise, occlusion free. Features are extracted from these images like Hog, edge detector; discriminative patches which are used for recognition. Shechtman et al. [37] used the correlation features to find the matching between two intensity patterns of images. Higher the correlation between the sequences implies that they belong to each other. The correlation between the two images can be found by matching the test image with the whole data set library. This method is useful in the representation of complex activities also but the time computation is large. Discrete Fourier transform (DFT) gives the global features of the image in the frequency domain. Kumari and Mitra [38] compute the DFT values for the silhouette images which are used as a feature vector for activity recognition as the feature vector computed on the image blocks for activity recognition. They assumed that background and foreground object intensity values are different, therefore their DFT values may also vary. But this method limited to single human detection and view point dependent. Liu et al [39] presented the action recognition based on the Pyramidal Motion feature (PMF). PMF is extracted from the optical flow algorithm based on Lucas Kanade applied to each frame of the video sequence. An Adaboost learning algorithm is used for selection of frames having most discriminating features and finally SVM classifier is used for classification. This method is simple and robust as it does not require any vocabulary formation neither interest point representation.

Chapter 3

Human Structure Representation

In this chapter, we present the 2D posture representation of actions based on Spatio-temporal methods. Through the posture representation, we can easily understand the human motion phenomena. We mainly focus on the structure representation which is based on the template based models. Further, we explain the feature extraction techniques that are employed on the static posture images. We also see that the features we used here are strong enough to discriminate the different actions.

3.1 Formation of MHI/MEI Images

The concept of motion history images (MHI) and motion energy images (MEI) is introduced by Bobick and Davis [27]. They represent the motion information of objects in the video sequence through the static images. They calculate the '*Hu*' moments on these images which they used as feature vectors for classification. It is one of the popular methods in action recognition and also robust in structure representation. The MHI images given the information about the motion where MEI images represent what kind of motion and in which direction. We worked on the same posture representation idea which is further embedded with the motion feature model. MHI/MEI images give the global representation of human posture.

MEI images are described as the sum of a binary image sequence which represents the region of motion. These MEI images are more stable than the MHI images as they give the hollow representation of the action [40].

Let $D(x, y, t)$be a binary image sequence which can be obtained from subtracting the image from the background and threshold process will convert into a binary image. The binary MEI $E_\tau(x, y, t)$ (τ is the duration) is defined as [42]:

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

MHI's images represent the latest information of action with brighter parts in grayscale images. Motion history images (MHI) $H_\tau(x, y, t)$ defined by Bobick [27] represented as:

$$H_\tau(x,y,t) = \begin{cases} \tau & if D(x,y,t) = 1 \\ \max(0, H_\tau(x,y,t-1) - 1) & otherwise \end{cases} \quad (2)$$

There is also another method of generating these images on the basis of frame differencing, which is explained as [41].

Let $I(x,y,t)$ be any image sequence and $I(x,y,t+1)$ be the next frame of previous frame,

$$D(x,y,t) = |I(x,y,t+1) - I(x,y,t)| \quad (3)$$

$D(x,y,t)$ will threshold to generate the Binary image;

$$B(x,y,t) = \begin{cases} 1 & D(x,y,t) \geq \theta \\ 0 & D(x,y,t) < \theta \end{cases} \quad (4)$$

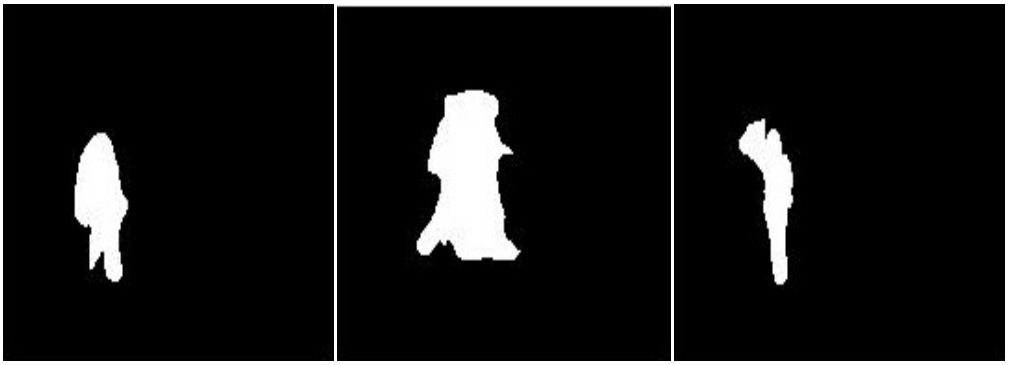
'θ'will be the threshold parameter and for MEI we take the sum of Binary images.

Then MHI is given as

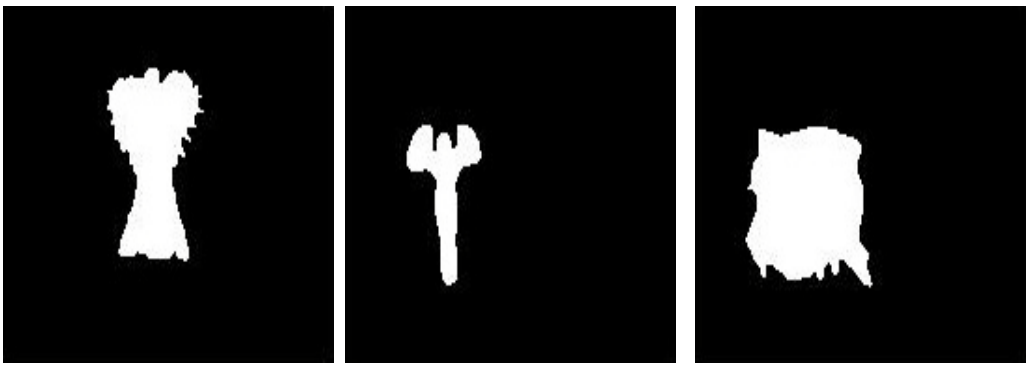
$$H_t(x,y) = max\{\cup_{t=1}^{N-1} B(x,y,t) \times t\} \quad (5)$$

In our method, we choose the $'\tau'$ parameter of finite interval after which the motion does not starts repeating. Like in case of 'Running' and 'Walking' we choose the $'\tau'$ parameter on the foot forward to another. This will give us an easy view representation and does not overwrite the information. But we have to choose the value of $'\tau'$ manually because the interval length changes for each action. For the generation of MHI images we use the frame differencing method as it is simple and does not involve the background subtraction method.

MEI Images



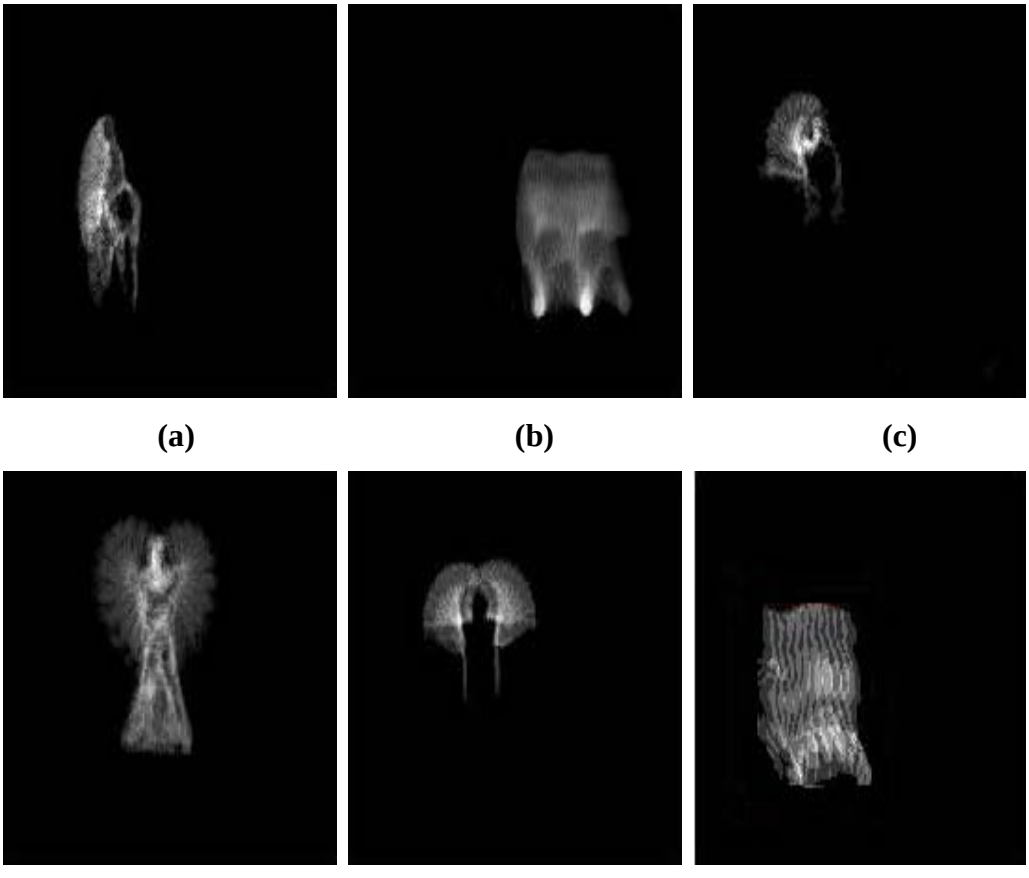
(a) **(b)** **(c)**



(c) (d) (e)

Fig 3.1: Representation of MEI images of (a) bend (b) walk (c) one hand wave (d) jump_jack (e) two hand wave (f) run activity on Weizmann data set.

MHI Images



(a) (b) (c)

(c) (d) (e)

Fig 3.2: Representation of MHI images of (a) bend (b) walk (c) one hand wave (d) jump_jack (e) two hand wave (f) run activity on Weizmann data set.

In representation of MHI/MEI images, we have some limitation. They will give only the information about the happening of action. The features extracted from these images are not sufficient to discriminate the various actions. The images are view

dependent and time 'τ' dependent. With the increase in value of$'\tau'$, the motion parameter information starts loosing because of overwriting of pattern of movements of actions like periodic movement of 'bending' action, with every periodic motion it will overwrite on the previous information. These images are successful in discrimination the actions like 'Bending', ' Running' but it is not effective in case of ' Running' and 'Walking' which follows the same patterns of movement.

3.2 Action Energy Image

V. Chandrashekhar et al. [43] introduced the concept of Action energy images (AEI) in his work for human action recognition where he formed the average Binary images which are more robust in the representation of MHI/MEI images because manually time dependency parameter has removed. Bhanu et al. [44] also introduced the Gait energy image (GEI) concept which is similar to AEI also. In case of AEI images higher intensity values represent the static motion whereas motion in action images represented by the low intensity values. It resolves the variation in the representation of 'running' and 'walking' problem because intensity variation will show slow and fast variation in motion. We further used the similar concept of Action Energy Images (AEI) for the representation of actions which is same concept related to Average Motion Energy image or Gait Energy image. An AEI image provides us most dynamic information in 2-dimensional static action images.

The equation for the Action Energy Image is defined as [1]

$$D(x,y,t) = \frac{1}{N}\sum_{t=1}^{N} B(x,y,t) \quad (6)$$

Where 'N' is represented as frame number and $B\ (x,y,t)$ is the Binary Silhouette in the sequence, 't' is the time instant of fame, 'x' and 'y' shows the pixel value.

In our work we choose the $'t'$ parameter equal to frame length.

Formation of AEI Images

For the formation of AEI images first we segment out the object using background subtraction method. Then the appropriate threshold parameter is set to obtain the Binary Silhouette image. We use a median filter in the pre-processing which will eliminate the small noises. Basic morphological operations are also employed such erosion, dilation. These operations help in separating the individual elements and join

the similar structure elements in an image. The obtained Silhouettes are not of same sizes therefore the normalization process is also performed.

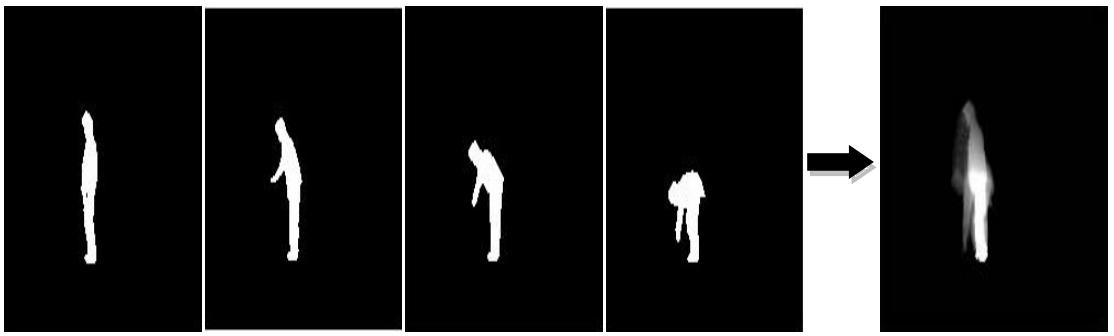
Extraction of Silhouette

(a) Foreground Image **(b) Subtracted Image** **(c) Silhouette Image**

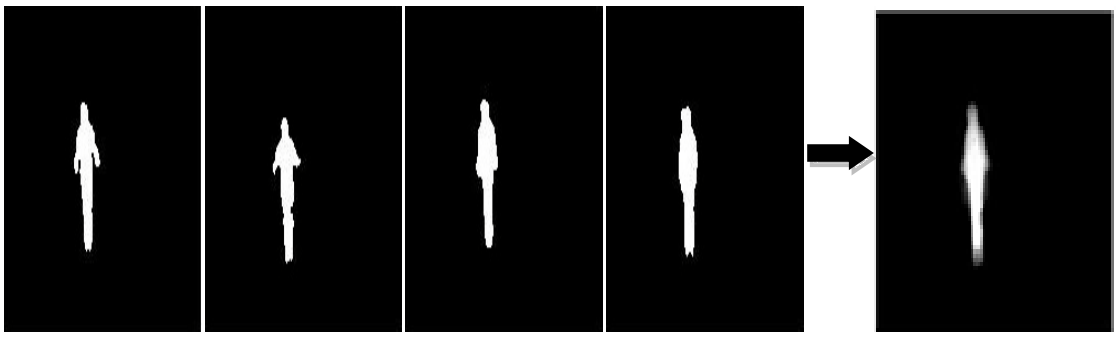
Fig 3.3: Representation of Silhouette image from Background Subtraction method

After forming the silhouette images, we divide the binary images by the frame length and obtain the AEI images. We show some of the representation of AEI images.

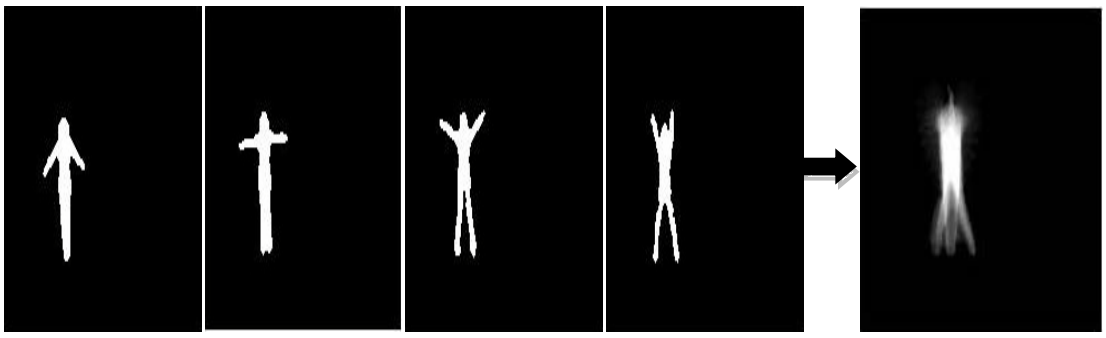
AEI Images



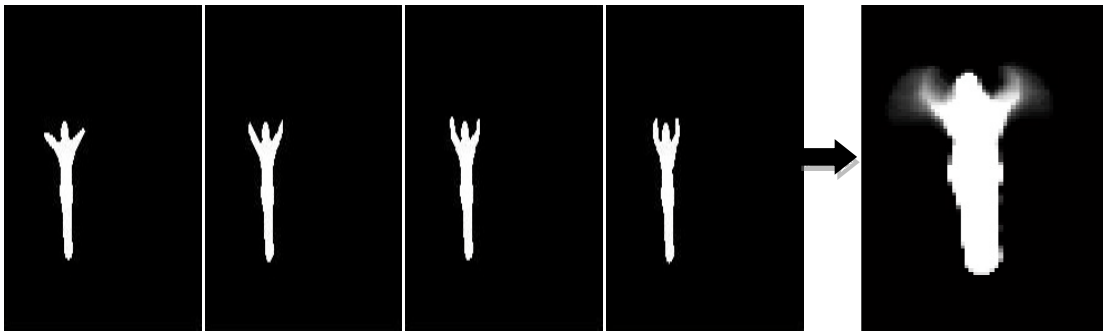
(a) Some of the Binary Silhouette Frames of Bending action and their AEI image



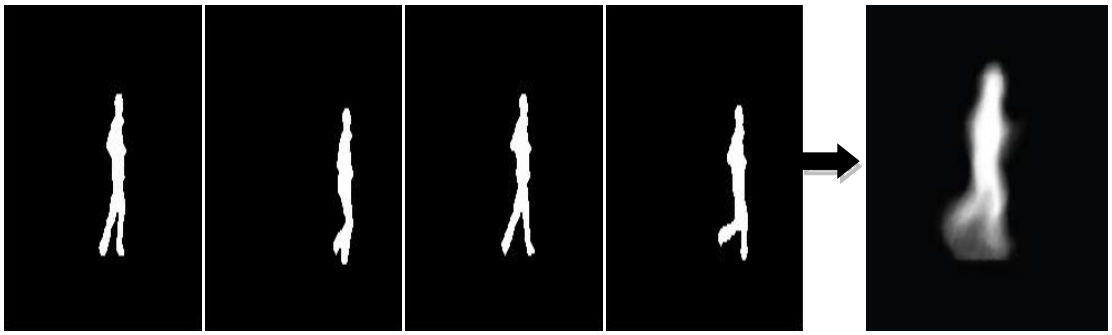
(b) Some of the Binary Silhouette Frames of Jump in place action and their AEI image



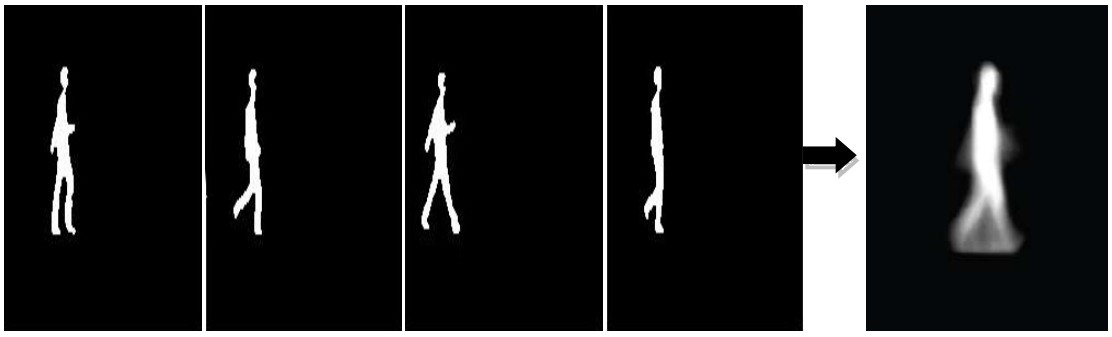
(c) Some of the Binary Silhouette Frames of Jump jack action and their AEI image



(d) Some of the Binary Silhouette Frames of Two Hand waving action and their AEI image



(e) Some of the Binary Silhouette Frames of running action and their AEIimage



(f) Some of the Binary Silhouette Frames of Walking action and their AEI image

Fig 3.4: shows some of the silhouette frames of the activity. These Binary frames sum up and averaged to give the AEI images. (a) Bending activity (b) Jumping in place, (c) Two hand wave (d) Jumping jack, (e) Run, (d) Walk.

3.3 Feature Extraction Methods

We apply the local descriptor Pyramid of Histogram (PHOG) on these images that will describe the shape of posture in an efficient manner. Besides the PHOG we find the directional variation in pixels which will give more internal information about the image. Local descriptor PHOG is performed on region of interest images (ROI) which will give more localized and noise free information. ROI will also help in making these images invariants to scale and translational variations.

3.3.1 Spatial shape descriptor – PHOG

The concept of the Pyramid of histogram (PHOG) was introduced by the A. Bosch et al. [45] in object classification, but now it's widely used in the face recognition or human shape representation and classification. It is based on the concept of Histogram of Oriented Gradient (HOG) [10] but compared to HOG it not only represents the object shape, but also gives the spatial information which will used as the discriminating feature for the representation of different shape. HOG describes the shape on the basis of the orientation of gradients in a particular region, but in case of PHOG each region is further divided into sub-regions and orientation of edges are counted in finer scale [46]. Therefore, we can say that PHOG works locally for shape representation. We explain the concept of PHOG by simple algorithm which we are using on posture models explained in the previous section.

ALGORITHM FOR PHOG

We compute the PHOG vector on the ROI image therefore we first extract the ROI by forming the bounding box around the action image.

Step 1: We extract edge contour of MHI/MEI/AEI using the canny edge detector, which is used to describe shapes.

Step 2: Segmentation of Image: - Divide the image into sub-regions to compute the histogram of the gradient vector. First level is the entire image, in the second level, the image is segmented into 4 sub-regions and in third level previous sub-regions are further divided into four smaller sub-regions.

Step 3: Find the HOG vector in each sub-region at each pyramidal resolution level. The magnitude $M\ (x, y)$ and orientation $\theta\ (x, y)$ of the gradient at any point (x, y) are calculated as follows:

$$M\ (x, y) = \sqrt{\mathrm{G_X(x,y)^2\ +\ G_y(x,y)^2}} \tag{7}$$

$$\theta(x, y) = arctan\frac{G_x(x,y)}{G_y(x,y)} \tag{8}$$

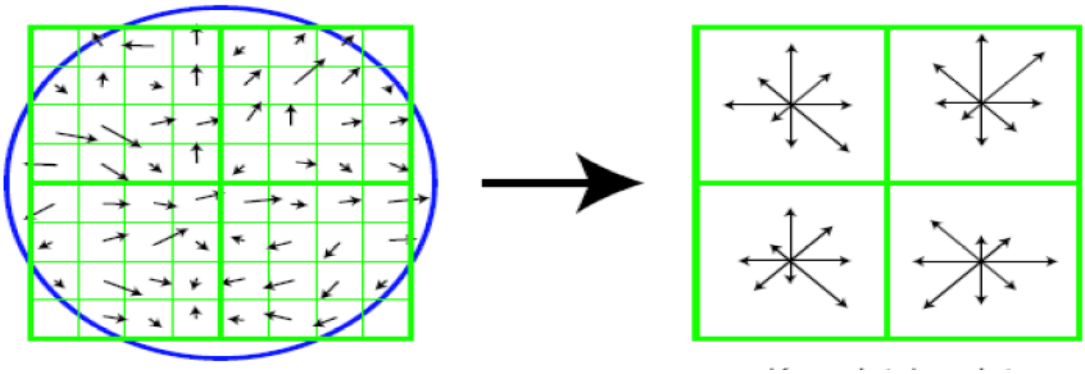
Where $G_x(x, y)$ and $G_y(x, y)$ are image gradients along the x and y directions.

Step 4: Gradient Edge Orientation is computed for each pixel. Each sub-region is quantized into the 8 orientation bins, evenly spaced over 0^o-180^oor 0^o-360^o.

Step 5: To find out the key point descriptor in the bin, voting method is used. The higher the magnitude of pixel representation in bin refers to the key point descriptor. Each bin represents the edge orientations.

Step 6: At last, we combined the HOG vectors at different pyramid resolution levels to form a final PHOG vector with dimension of d =$K\sum_{l=0}^{L} 4^l$ to represent the whole ROI.

Step 7: In our method we divide the image into two levels (L = 2) and HOG vector was quantized into 8 orientation bins in the range of [0, 360]. Therefore, at L=2 and k=8 final HOG vector is a 168 vector (8× [4^0+4^1+4^2]).

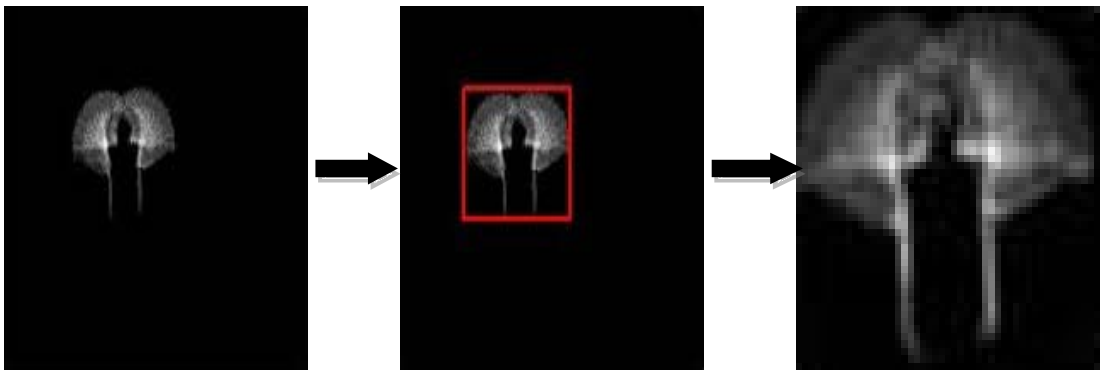


(a) (b)

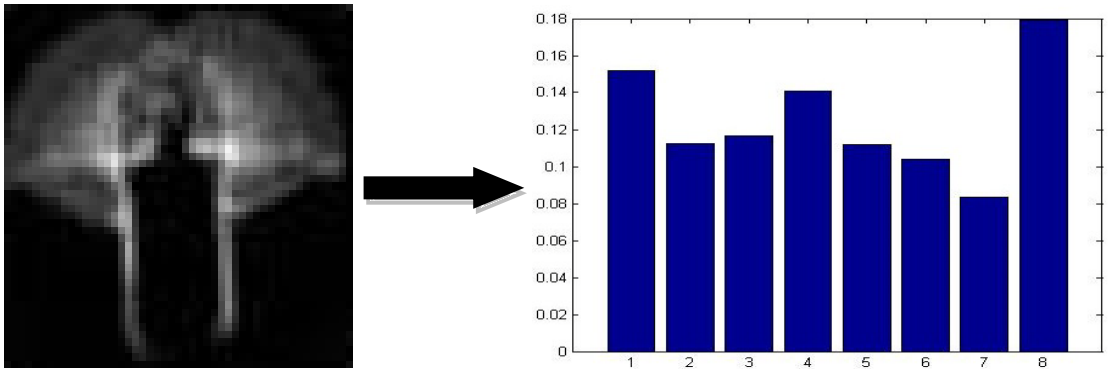
Fig 3.5: Formation of the gradient key point descriptor. (a) Each region is divided into four sub regions. This is obtained by first selecting the 8 ×8 window and then divided into 4 sub parts. Each bin shows the image gradients in different directions. (b) Represent the Key point descriptors in 4×4 cell with higher magnitude angle [10].

Representation of Algorithm

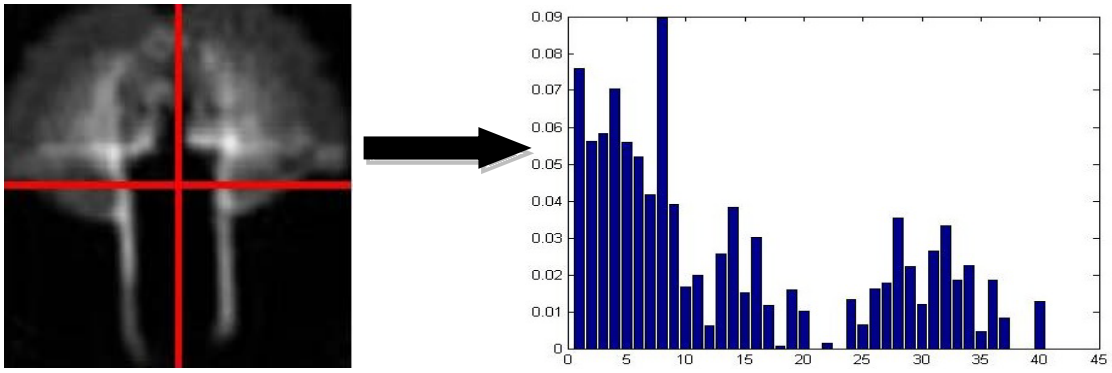
First, we show the selection of ROI from the MHI image of two hands waving action and then we represent the computation of PHOG vector at different levels.



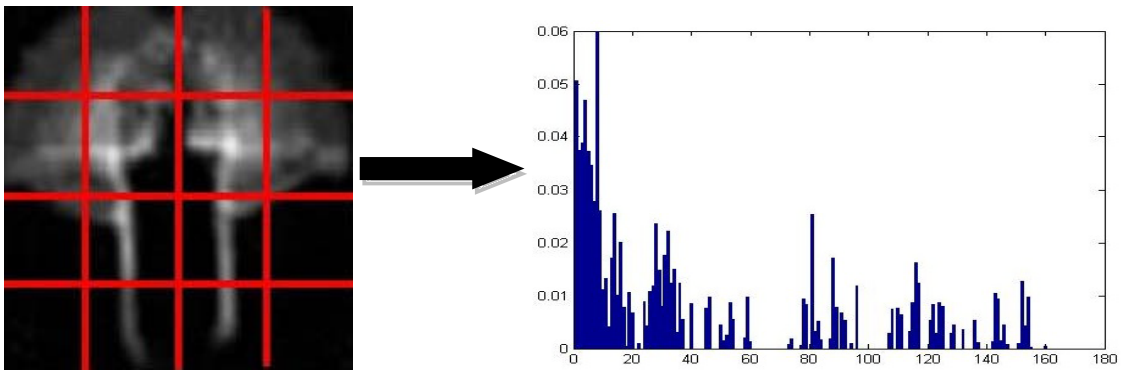
(a) Original image **(b) ROI selection** **(c) Normalized image**



(d) PHOG vector computed at level 0. At level 0, the ROI would be entire image. The x axis and y axis represents the number of bins and magnitude respectively.



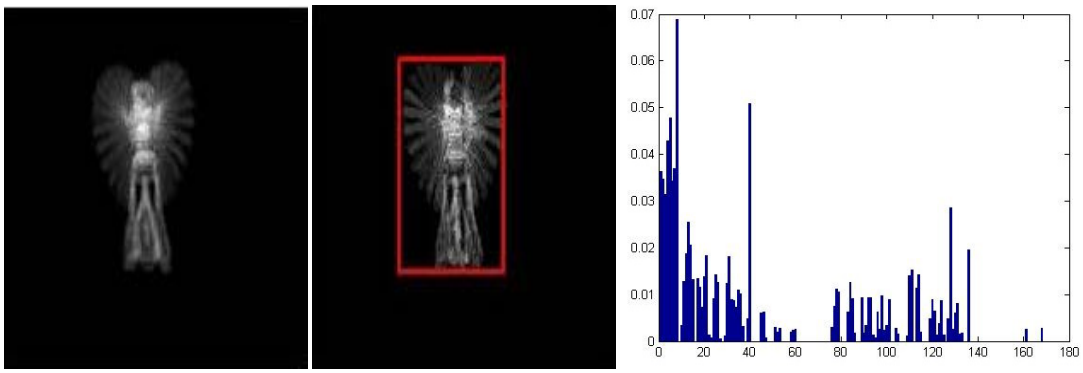
(e) PHOG vector computed at level 1. At level 1, ROI divided into 4 sub-regions. The x axis and y axis represents the number of bins and magnitude respectively.



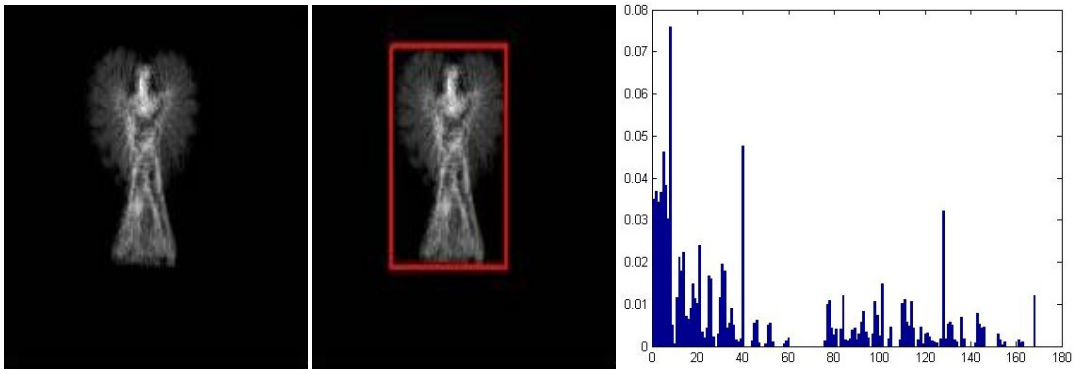
(f) PHOG vector computed at level 2. At level 2, ROI divided into 16 sub-regions. The x axis and y axis represents the number of bins and magnitude respectively.

Fig 3.6: Overview of the computation of PHOG vector.

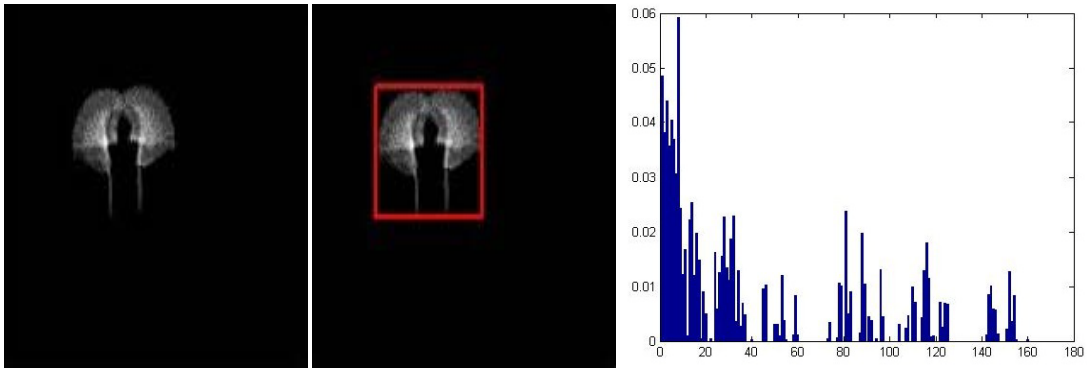
PHOG Computation of Different and Similar MHI Images



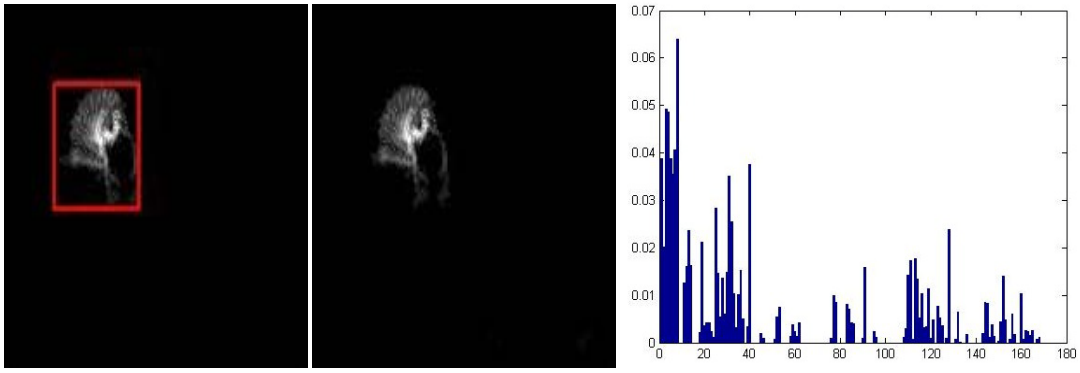
(a) Jumping Jack (1)



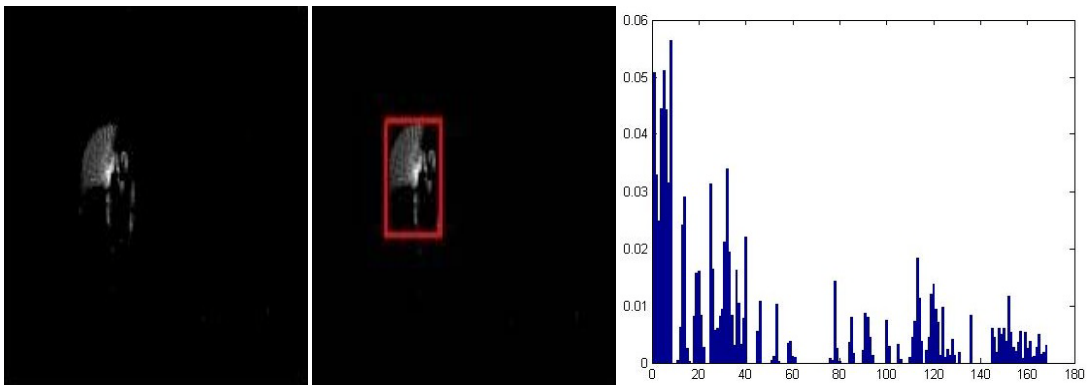
(b) Jumping Jack (2)



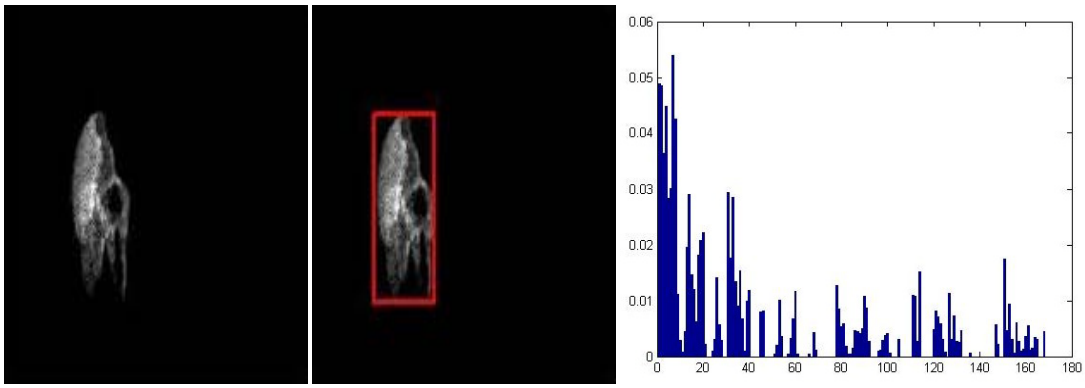
(c) Two Hand Wave



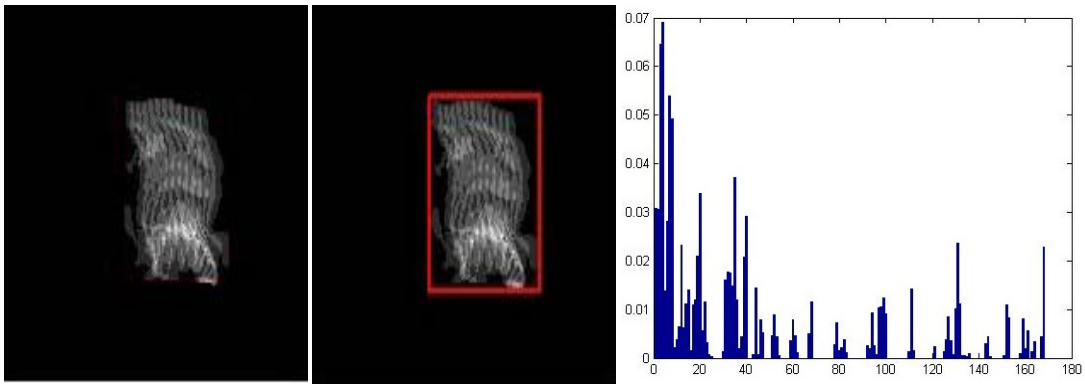
(d) One Hand Wave (1)



(e) One Hand Wave (2)



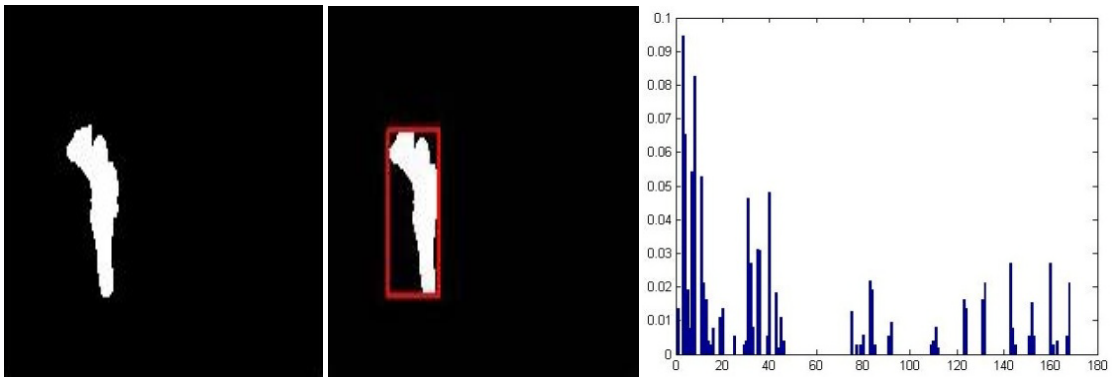
(f) Bending



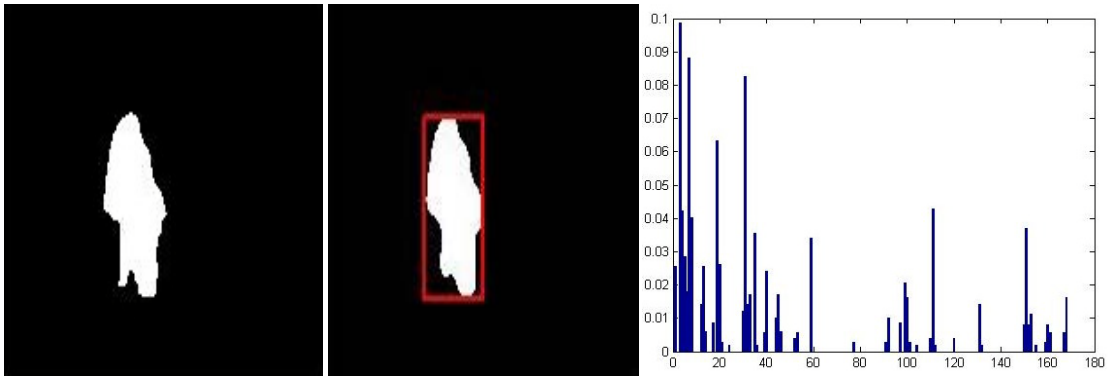
(g) Running

Fig 3.7: Spatial Pyramid Histogram representation of jumping jack, two hand wave, one hand wave, bending of same and different categories at Level2 decomposition. First, we show the MHI images of different actions. From these images we select the region of interest (ROI) by forming the bounding box. We apply the PHOG vector to these ROI images. The level of decomposition is at two levels. From the diagrams (a), (b) and (d), (e) it is clear that PHOG representation is similar for two similar activities performed by different people. The magnitude of the peaksis slightly varied but the representation of histograms are same. From (c), (f), (g) PHOG representation shows that it is discriminating between different categories.

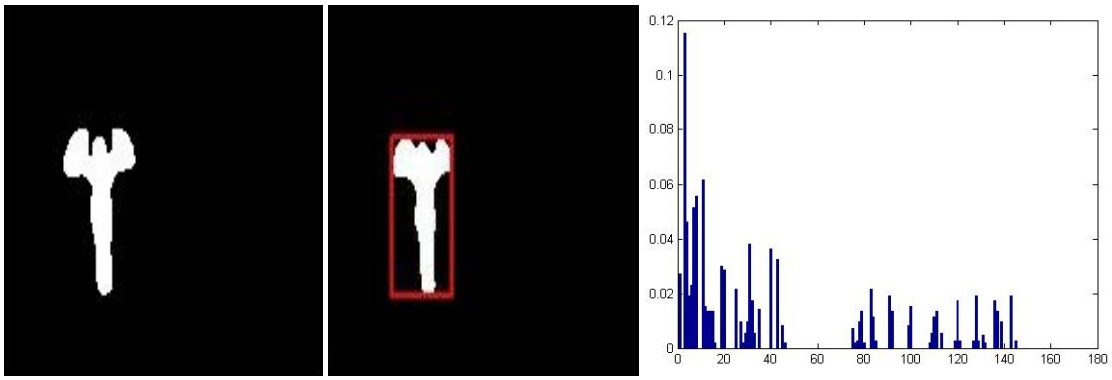
PHOG Computation of Different MEI Images



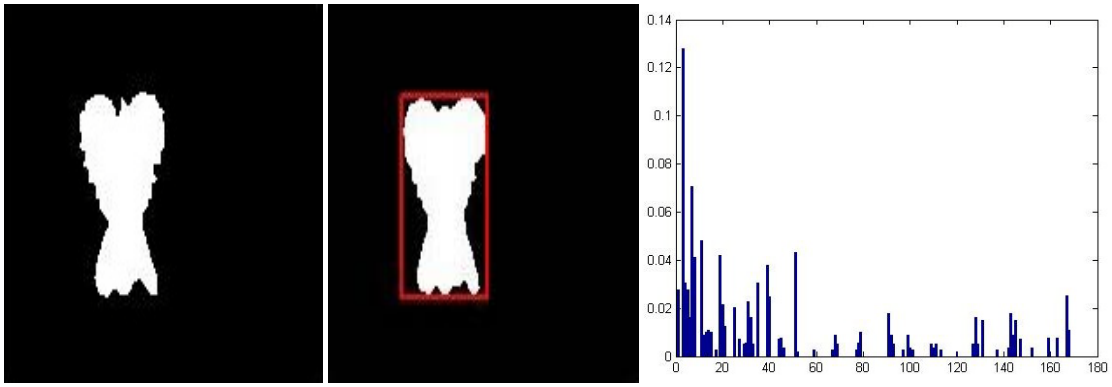
(a) One Hand wave



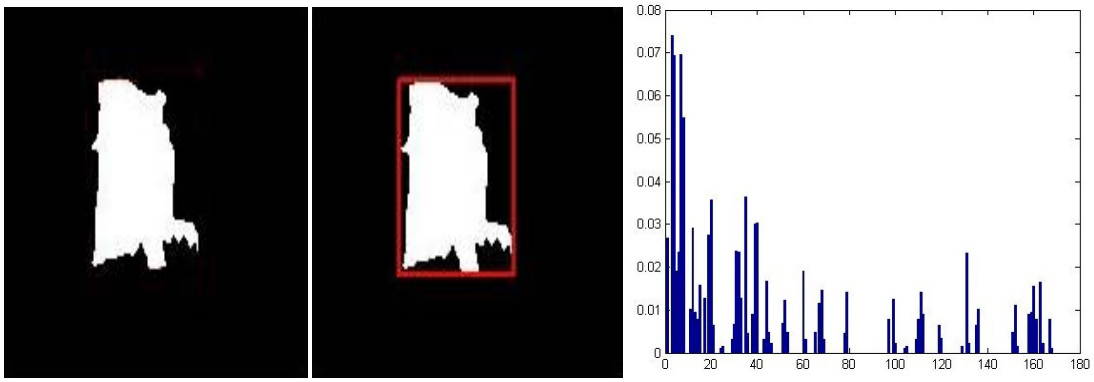
(b) Bending



(c) Two Hand wave



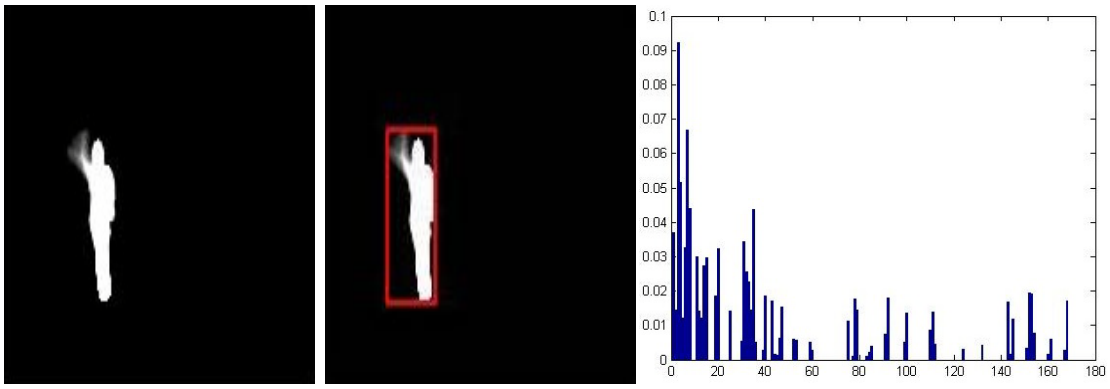
(d) Jumping Jack



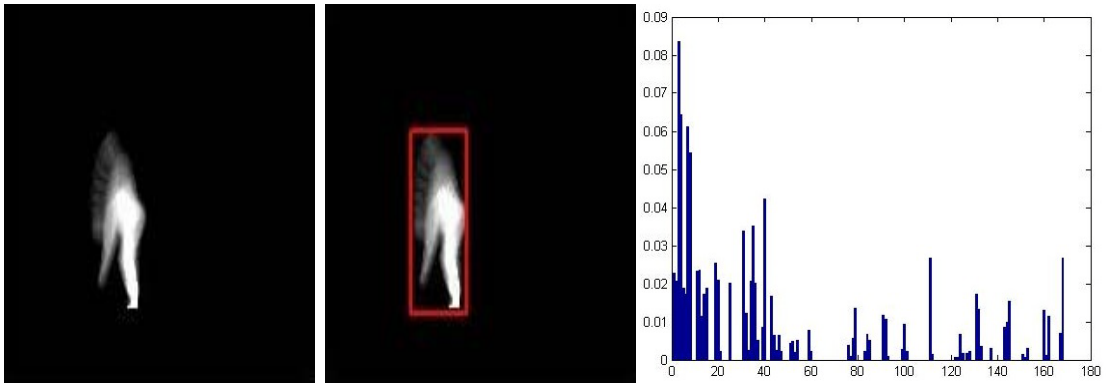
(e) Running

Fig 3.8: Pyramid Histogram representation of MEI images of one hand wave, bending, two hand wave, jumping jack, activities. We apply the PHOG vector to these ROI images. The level of decomposition is at two levels. Representation clearly shows that for different actions PHOG representation is different.

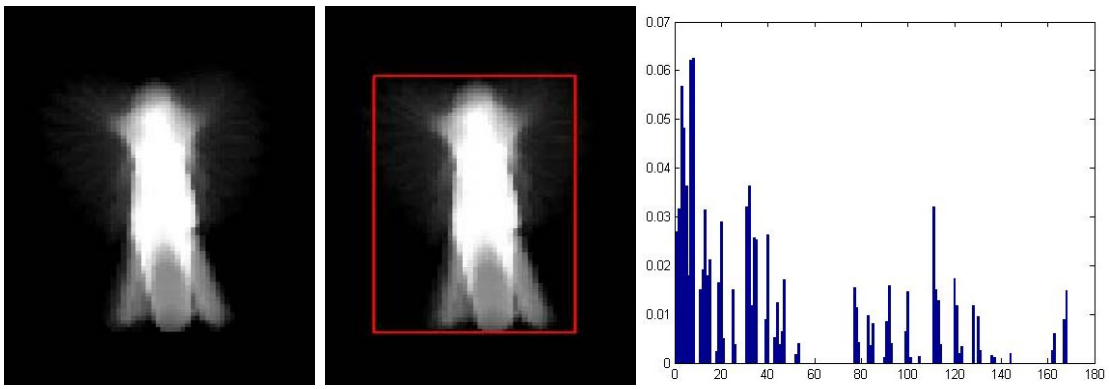
PHOG Computation of Different AEI Images



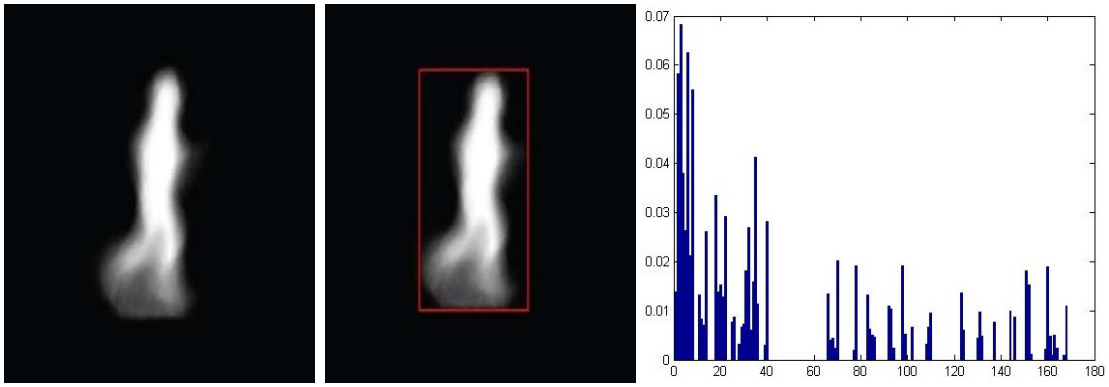
(a) One Hand wave



(b) Bending



(c) Jumping Jack



(d) Running

Fig 3.9: Pyramid Histogram representation of one hand wave, bending, jumping jack,running activities at Level2 decomposition.

The above image representation shown that PHOG vector gives discriminating feature vectors because similar action have nearly same spatial distribution and different action represents different pyramid distribution. But the application of PHOG will work more efficiently if the shape model representation is accurate and extraction of ROI performed correctly.

3.3.2 Directional Pixels along X-Y direction.

We find the pixel value of the shape representing the action along x-y directions. These values give us more information about the shape. The variation in the values gives us the idea about the action. We normalize the values obtain in x-y direction to reduce the redundant information. From these values we obtain the mean as our feature vector for the recognition process. The equation for computing the pixels along the x-y direction described below [47, 48];

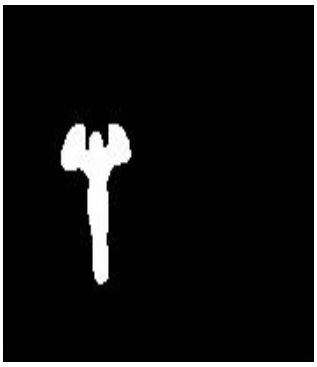
$$H_k(x) = \sum_{l=0}^{N-1} \frac{A\,(k,l)}{\max\,((k))}) \quad , \qquad \text{k= 0, 1..... M-1} \qquad (9)$$

$$V_l(y) = \sum_{l=0}^{M-1} \frac{A(k,l)}{\max[A(l)]} \quad , \qquad l = 0, 1.....N-1 \qquad (10)$$

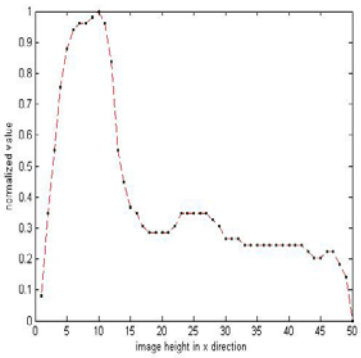
Height and Width represented by 'M' and 'N' of the action image. Mean value represented in x and y direction

$$H_X = \frac{1}{M}\sum_{k=1}^{M} H_k(x) \quad V_y = \frac{1}{N}\sum_{l=1}^{N} V_l(y)$$

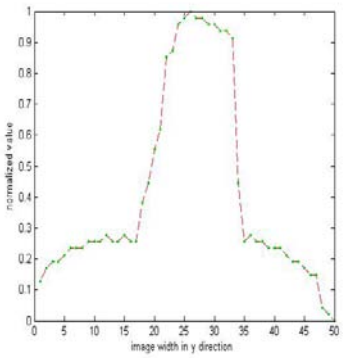
Pixels Representation



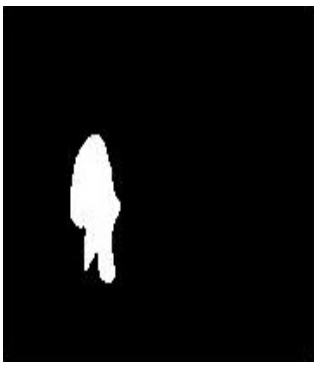
(a) Two hand wave MEI



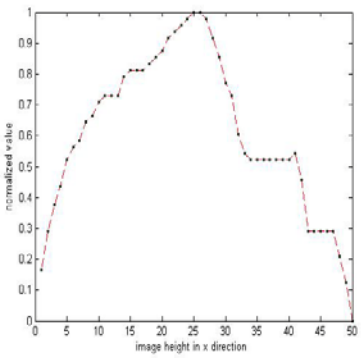
(a1) pixels in x-direction



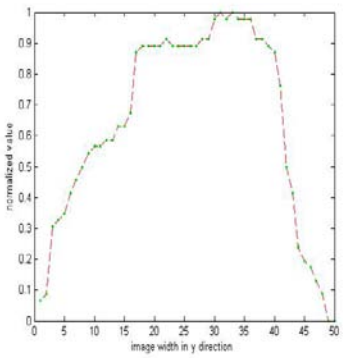
(a2) pixels in y-direction



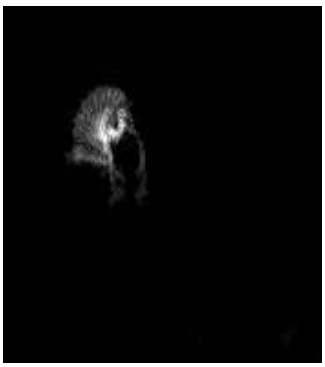
(b) Bend MEI



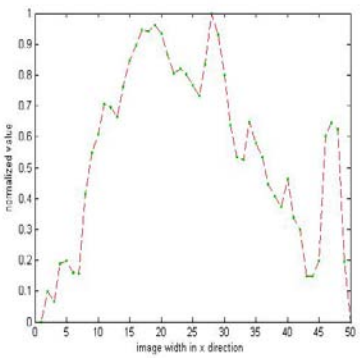
(b1) pixels in x-direction



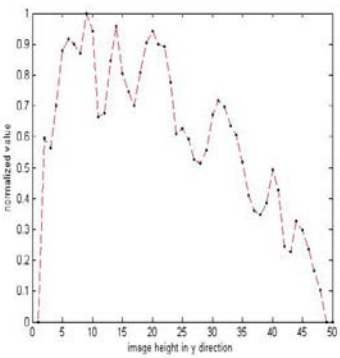
(b2) pixels in y-direction



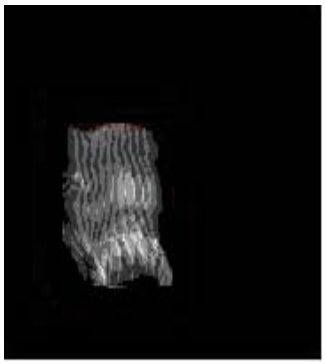
(d) One hand wave MHI

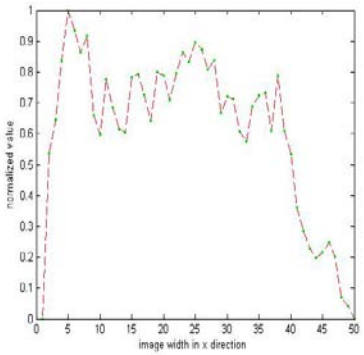


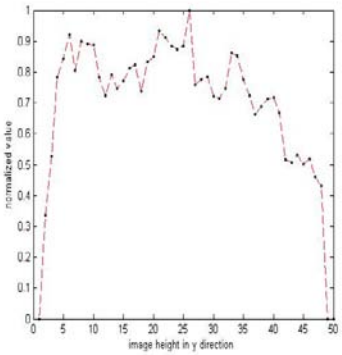
(d1) pixels in x-direction



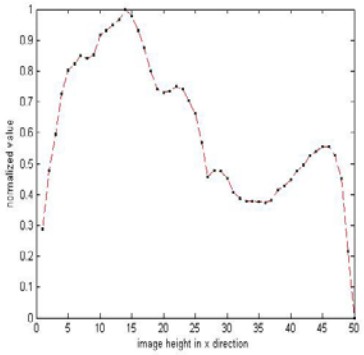
(d2) pixels in y-direction

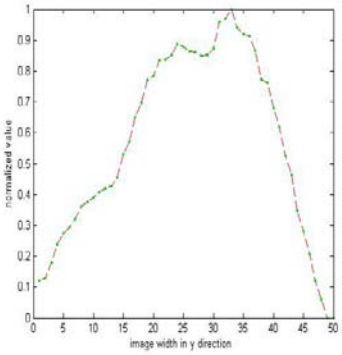






(e) Running MHI **(e1) pixels in x-direction** **(e2) pixels in y-direction**





(f) Bend AEI **(f1) pixels in x-direction** **(f2) pixels in y-direction**

Fig 3.10: Above figure shows the pixel values along the x and y direction. Peaks shows that, the variation of pixel value is more compared the other areas. Consider (a) peak at the centre shows that more pixel are present as we go column wise (y-pixel) and as we go from x-axis it represents the presence of more pixels in upper portion. They are not totally giving the representation of the shape, but it only represents where pixels presence is more.

3.3.3 Discrete Fourier Transform (DFT)

We compute the 2-DFT on the pixel values that we calculated in the x-y direction. The Fourier representation will transform the features in the frequency domain. The 2-D Fourier transform represented as;

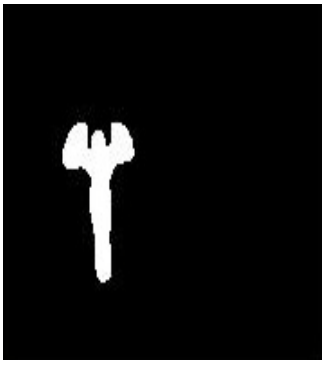
$$F(u,v) = \frac{1}{M \times N} \sum_{x=0}^{M} \sum_{y=0}^{N} f(x,y)\, e^{-j\,2\pi(\frac{ux}{M},\frac{vy}{N})} \quad (9)$$

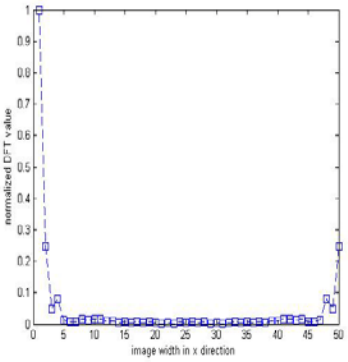
$$f(x,y) = \frac{1}{M \times N} \sum_{u=0}^{M} \sum_{v=0}^{N} F(u,v) e^{j\,2\pi(\frac{ux}{M},\frac{vy}{N})} \quad (10)$$

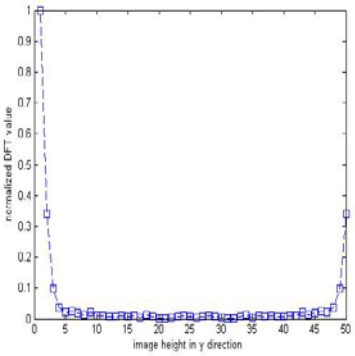
From the Fourier transform values we find the mean value which we used as a feature vector. In the end the combined feature vector from the shape information will be a vector consisting of PHOG, mean value pixels in x-y direction, mean values of 2-DFT.

Final shape Vector= $\left[(PHOG_1, PHOG_2 \ldots\ldots PHOG_n), H_x, V_y, F_x, F_l\right]$

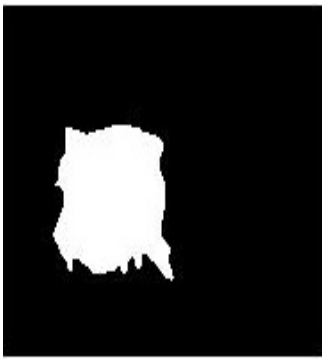
Fourier Representation

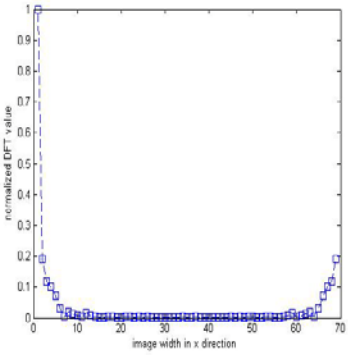


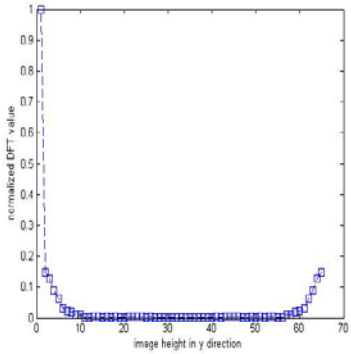




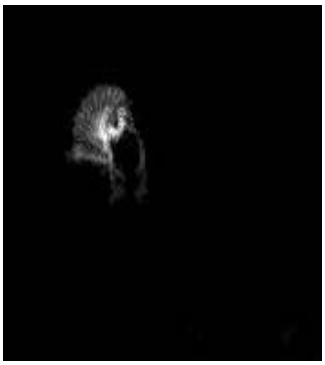
(a) Two hand wave MEI **(a1) Frequency in x-dir^n** **(a2) Frequency in y-dir^n**

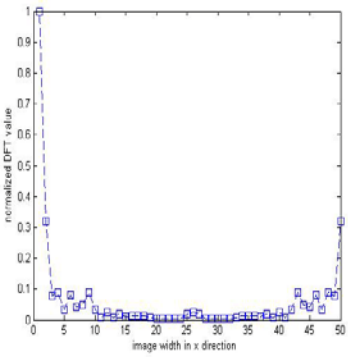


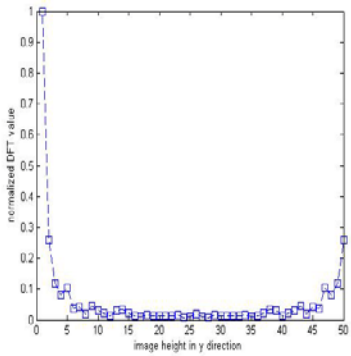




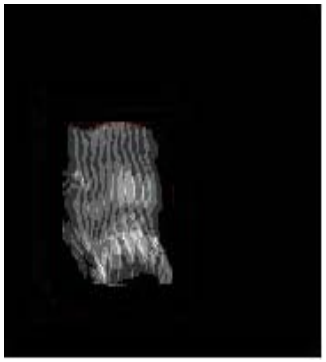
(b) Running MEI **(b1) Frequency in x-dir^n** **(b2) Frequency in y-dir^n**

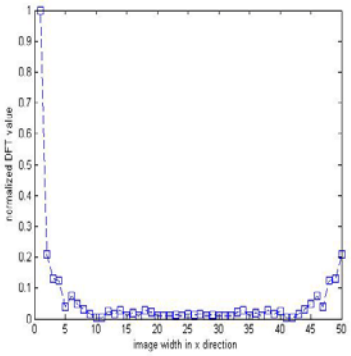


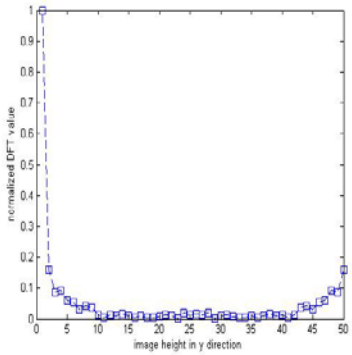




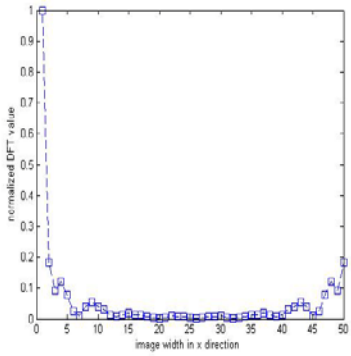
(c) One hand wave MHI **(c1) Frequency in x-dir^n** **(c2) Frequency in y-dir^n**

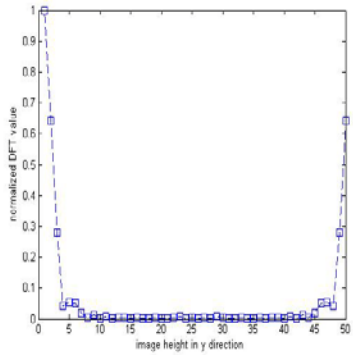




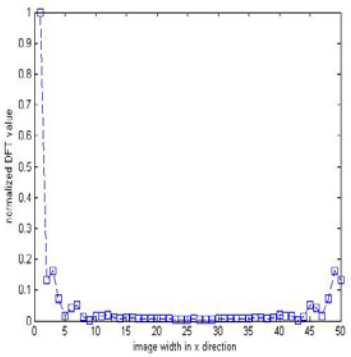


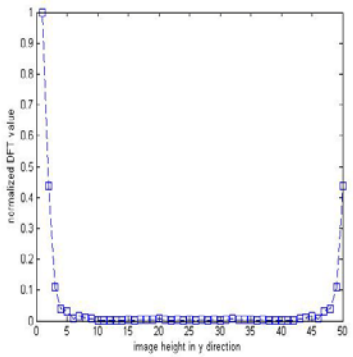
(d) Running MHI **(d1) Frequency in x-dir^n** **(d2) Frequency in y-dir^n**





(e) Two hand wave AEI **(e1) Frequency in x-dir^n** **(e2) Frequency in y-dir^n**





(f) Running AEI **(f1) Frequency in x-dir^n** **(f2) Frequency in y-dir^n**

Fig 3.11: Representation fourier transform in x- y direction.

Chapter 4

Human Body Motion Descriptor

In our work we used the R-transform to represent the motion features which are obtained from Radon Transform. For an image, Radon transform (RT) gives the projection of lines in all directions. We summed all the pixel values in the direction of lines which are described as by Radon coefficients. These coefficients changes with translation, rotation and scaling of an object. R-transform shows the invariance to translation and scaling of an object while it changes with the rotation of an object. We used this idea in the representation of human action where arms, legs are continuously changing position with time, which give us satisfactory discriminating features in representation of actions. We first discuss the previous work related to recognition of actions based on Radon transform followed bya brief introduction about Radon transform properties and its improved version R-transform.

4.1 Related Work

Singh et al. [49] introduced the unique method which makes use of Radon Transform (RT) for recognition of human pose. Human pose is represented by medial skeleton images of arms and hand direction. RT is applied to these skeleton lines for the detection of orientation of these lines. They used SMM (spatial maxima mapping) algorithm to find the matching between training and unknown images. Wang et al. [50] directly employed the radon transform on Binary silhouette images. Principal component analysis is used for the dimension reduction of the feature vectors. The hidden Markov model is used to design the different activities. They compared the RT with Zernike, Invariant, Wavelet and Pseudo Zernike moment descriptors and showed that RT performs better than others descriptors. Boulgouris et al. [51] used the Radon transform in the gait recognition model. In this method template is constructed from Radon silhouette images. Linear discriminant analysis (LDA) is further employed on a template which gives discriminating features. Li et al. [52] proposed the geometric transformation method used to represent the shape and appearance information of the action. For the representation of geometric transformation they used the Radon, Trace transform and image warping. Zhang et al. [53] gives the simple and effective approach of recognition of human activities through key frames. R-transform is computed on the

key silhouette frames extracted from the video cycle. Moustakas et al. [54] gives another method where he used RT with soft biometrics for recognition of human gait. They used the biometric traits to represent the information about gait images. Khan et al. [55] introduced the abnormality activity recognition system based on R-transform. Jalal et al. [56] represents the human action in depth silhouette. For each depth silhouette, Radon transform is employed, which give invariance to silhouette in terms of scaling and rotation.

4.2 Introduction of Radon Transform

In 1917, Johann Radon brought the concept of Radon Transform. It is firstly introduced in the tomography but later on its properties finds great application in human action recognition system. With the help of Radon transform, projection of image can be obtained in a set of lines at different angle($0 - 179^0$). Along these lines, Radon transform is defined as the integral of the function (f) from -∞ to ∞, denoted as

$$R_f(\rho,\theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\delta(\rho - x\cos\theta - y\sin\theta)\,\partial x\,\partial y \qquad (11)$$

f(x, y) is the 2-D image over which RT is calculated,

δ(.)is defined as the Dirac delta function which is zero everywhere except at the origin.

$$\delta(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{elsewhere} \end{cases} \qquad (12)$$

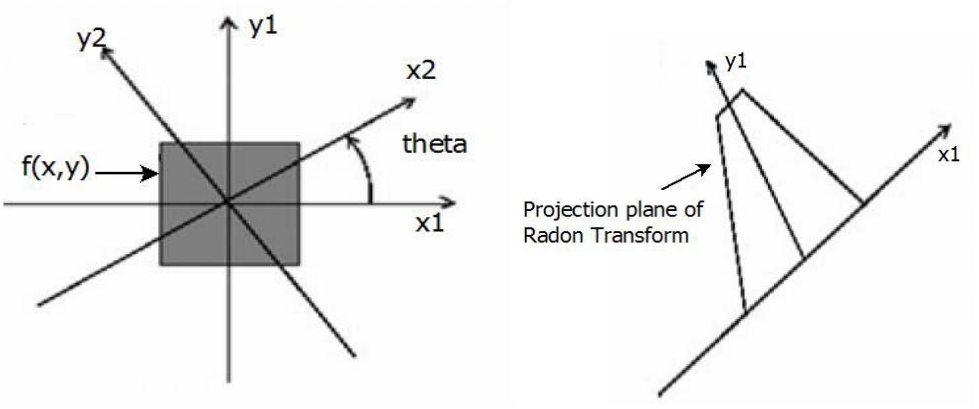
4.2.1Geometry of Radon Transform

The given below figure shows the single projection at a specified rotation angle [57]. To obtain the motion features we apply the RT on binary silhouette images. In the given below diagram, we present the view representation of radon transform in 2-dimensional coordinate system (x, y) of binary silhouette image in radon coefficients (ρ, θ).

The position of a point on the projection line is defined by ρ, expressed as

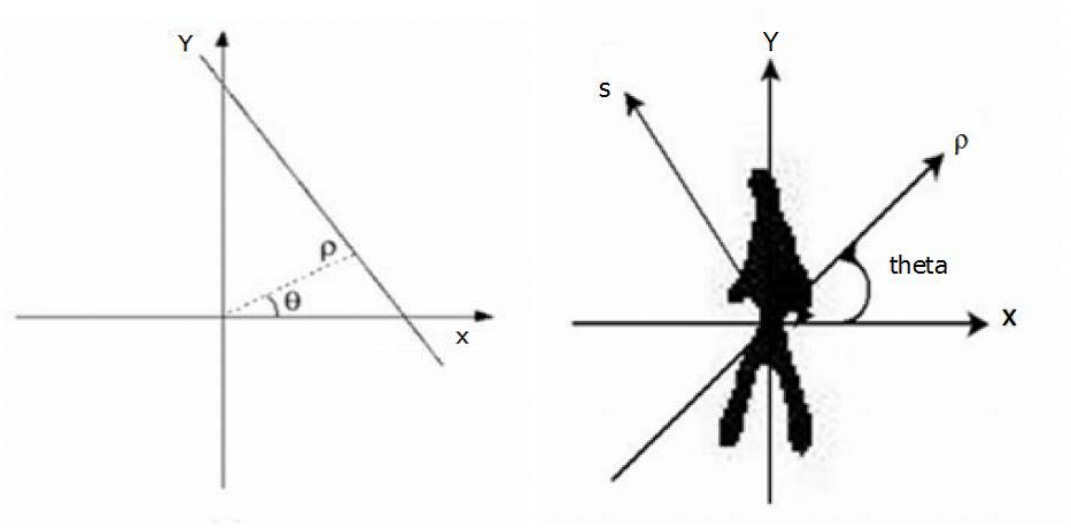
$$\rho = x\cos\theta + y\sin\theta (0 \le \theta \le \pi)(-\infty \le \rho \le \infty) \qquad (13)$$

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



(a) (b)

Fig 4.1: Illustrates the projection of lines over 2-D function f (x, y) and x1, y1 represents the 2-D plane and Radon transforms computes the line integrals over the projection plane with the rotating angle (θ). (b) Radon transform projection plane.



(a) (b)

Fig 4.2: (a) Shows the angular (θ) and position (ρ) parameter. ρ and θ determine the position of the projection line along with radon transforms computes the integral summation is computed. (b) The ρ-s coordinate system. Here, θ is represented by theta.

4.2.2 Properties of RadonTranform

We define the properties of RT by using the concepts described by S. Tabbone [58]

1) **Superposition**

$$R_f(\rho,\theta)[f_1(x,y) + f_2(x,y)] = T_1(\rho,\theta) + T_2(\rho,\theta) \quad (14)$$

Where$R_f(\rho,\theta)f_1(x,y) = T_1(\rho,\theta)$ *and*

$$R_f(\rho,\theta)f_2(x,y) = T_2(\rho,\theta).$$

2) **Linearity**

$$R_f(\rho,\theta)[af(x,y)] = aT(\rho,\theta),\ a\text{is any rational number.} \qquad (15)$$

$$R_f(\rho,\theta)\ f(x,y) = T(\rho,\theta)$$

3) Scaling

If we scale by $'\alpha'$, then

$$[(R_f(\alpha\rho,\theta)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} R_f\ (\alpha(\rho - xcos\theta - ysin\theta),\theta)\partial x\partial y \qquad (16)$$

Put $\alpha(\rho - xcos\theta - ysin\theta) = v$, then

$$[\big(R_f(\alpha\rho,\theta)\big] = \frac{1}{\alpha^2}(R_f(\alpha\rho,\theta)$$

From the above equations it is clear that scaling parameter α results in scaling of both the magnitude and ρparameter.

4) **Translation**

If $\underset{a}{\rightarrow} = x_o, y_o$is a vector,

$$R_f(\rho,\theta) = R_f(\rho - x_o cos\theta - y_o sin\theta,\theta) \qquad (17)$$

Translation in the ρ coordinates by (x_o, y_o) will results the same amount of shifting in the initial ρ coordinates.

5) **Periodicity**

Periodicity is defined as

$$R_f(\rho,\theta) = R_f(\rho,\theta + 2k\pi),\ k \text{ is any integer value,} \qquad (18)$$

2π is the interval length.

6) **Symmetry**

$$R_f(\rho,\theta) = R_f(-\rho,\theta \pm \pi) \qquad (19)$$

7) **Rotation by θ_o**

If θ_1 is the rotation provided in the image, then it results shifting of θ_1 in angular parameter θ of Radon transform,

$$R_f^1(\rho,\theta) = R_f(\rho,\theta+\theta_1) \tag{20}$$

From the properties describe above it is clear that Radon transform can't restore all parameters of the original geometric transformation when translate, rotate or scale the image.

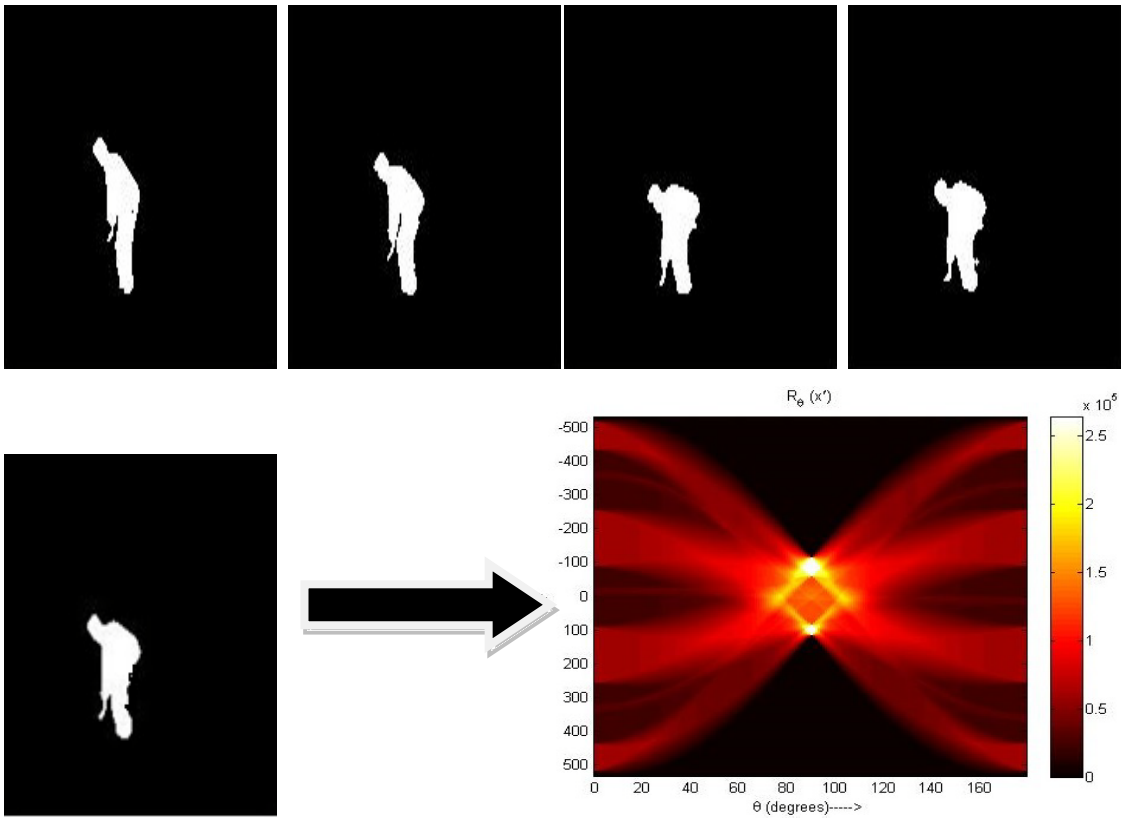


Fig 4.3: Radon Transform of Bending activity. Brighter portion denotes the summation of values along the projective lines.

Therefore S. Tabboneet al. [58] introduce the R- transform which is defined as Eq.

$$\mathrm{R}(\theta) = \int_{-\infty}^{\infty} \mathrm{R}_{\mathrm{f}}^2(\rho,\theta)\,\partial\rho \tag{21}$$

R(θ)is known as the R-transform which is an integral transform of the squared values of RT that are calculated at $0-179^0$angles [10,11]. In our work we used the same concept described in [7, 11] to represent the motion features. Compared to radon transform, r-transform only gives angular variation and remains invariant to

translational and scaling parameters. Radon transform gives the features in 2-D dimension (columns represent the different angles and rows represents the feature vector) while R-transform represents the features in 1-D dimension (column values are summed row wise gives reduced dimensions).

4.2.3 Properties of R-Transform

The properties of R-Transform are related to the Radon Transform. We describe these properties using earlier equations defined in RT.

1- Periodicity

$$R(\theta \pm \pi) = R(\theta)\,, \qquad \text{period is } \pi$$

Using Symmetry property of Radon Transform

$$R_f(\rho, \theta) = R_f(-\rho, \theta \pm \pi) \tag{22}$$

$$R(\theta) = \int_{-\infty}^{\infty} R_f^2(-\rho, \theta \pm \pi)\partial\rho$$

Put $-\rho = x$, then

$R(\theta) = \int_{-\infty}^{\infty} R_f^2(x, \theta \pm \pi)\partial\rho$, which is equal to $R(\theta)$

2- Translation

$$R(\theta) = \int_{-\infty}^{\infty} R_f^2((\rho - x_o cos\theta - y_o sin\theta), \theta)\partial\rho \tag{23}$$

Using Translation Property

If $\underset{a}{\rightarrow} = x_o$, yo is a vector, and

$$R_f(\rho, \theta) = R_f(\rho - x_o cos\theta - y_o sin\theta, \theta)$$

$$R(\theta) = \int_{-\infty}^{\infty} R_f^2((\rho - x_o cos\theta - y_o sin\theta), \theta)\partial\rho$$

Put $\rho - x_o cos\theta - y_o sin\theta = x$, then

$$\int_{-\infty}^{\infty} R_f^2(x, \theta)\partial\rho = R(\theta)$$

3- Rotation

$$R(\theta) = \int_{-\infty}^{\infty} R_f^2(\rho, \theta + \theta_O)\partial\rho$$

If $R_f^1(\rho,\theta) = R_f(\rho,\theta+\theta_o)$, Then

$$\int_{-\infty}^{\infty} R_f^2(\rho,\theta+\theta_0)\partial\rho = R_f^1(\theta+\theta_o) \quad (24)$$

4- <u>Scaling</u>

If we scale by α, then

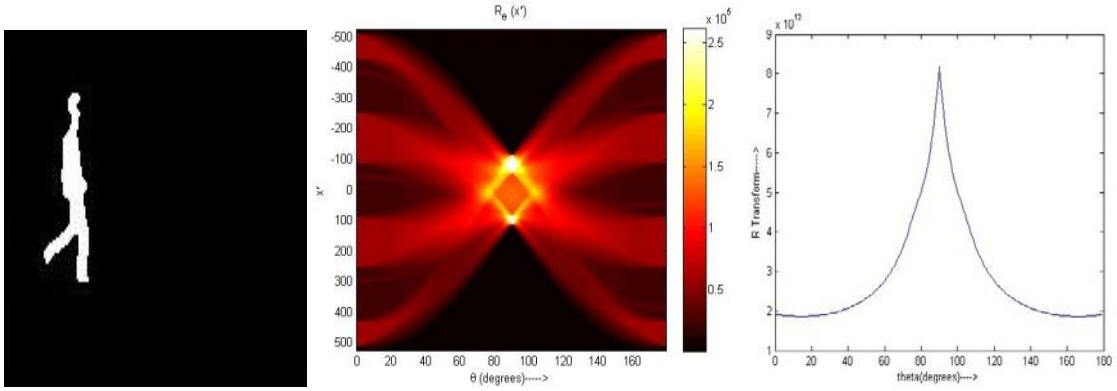
$$[(R_f(\alpha\rho,\theta)] = \frac{1}{\alpha^2}R_f(\alpha\rho,\theta) \qquad ,gives$$

$\int_{-\infty}^{\infty} R_f^2(\alpha\rho,\theta+\theta_0)\partial\rho$, Putting$\alpha\rho = x$ will resist

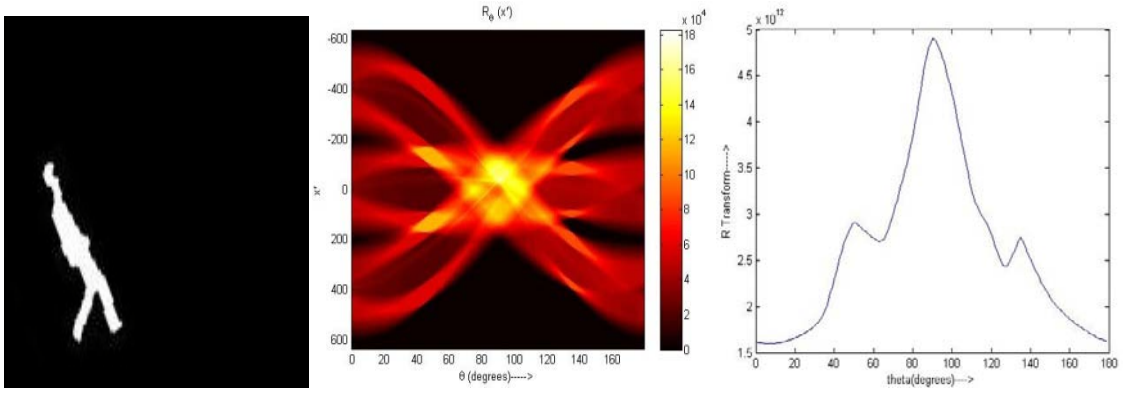
$$\frac{1}{\alpha^3}\int_{-\infty}^{\infty} R_f^2(x,\theta+\theta_0)\partial x = \frac{1}{\alpha^3}R(\theta) \quad (25)$$

From the above equation it is concluded that, $'\mathcal{R}'$ transform is invariant to translation in the plane, rotation in the plane will change the phase shift in $'\mathcal{R}'$ transform and scaling in original image does not show any effect on the representation of a signal but the magnitude will vary accordingly. Therefore R transform will be an efficient method in describing the shape where there is a change in rotation parameter such as when a person bends, walking and running casing where the angle is continuously changing during the movement of arms and legs. So it can be used as a feature vector in representing the motion in a discriminating manner. 1-D R- transform properties applied on the binary silhouette images showing the variation in case of rotation while remaining invariance to translate and scale variation.

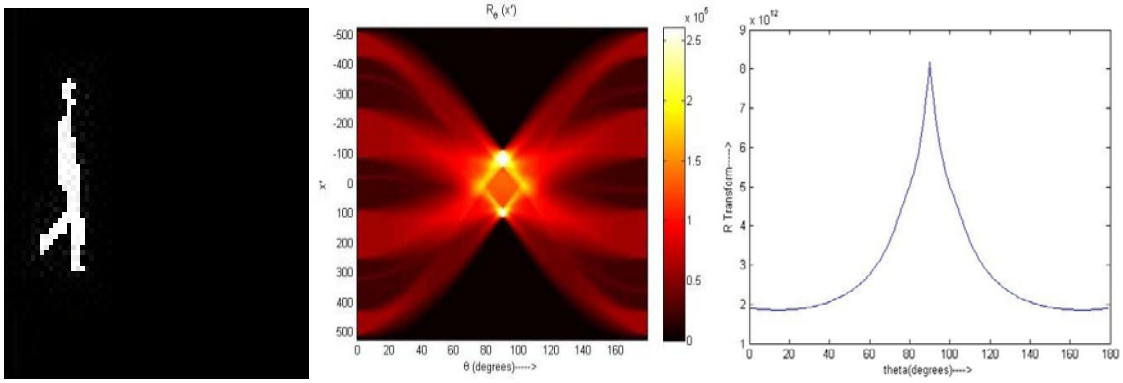
<u>Diagramatic view of Properties</u>



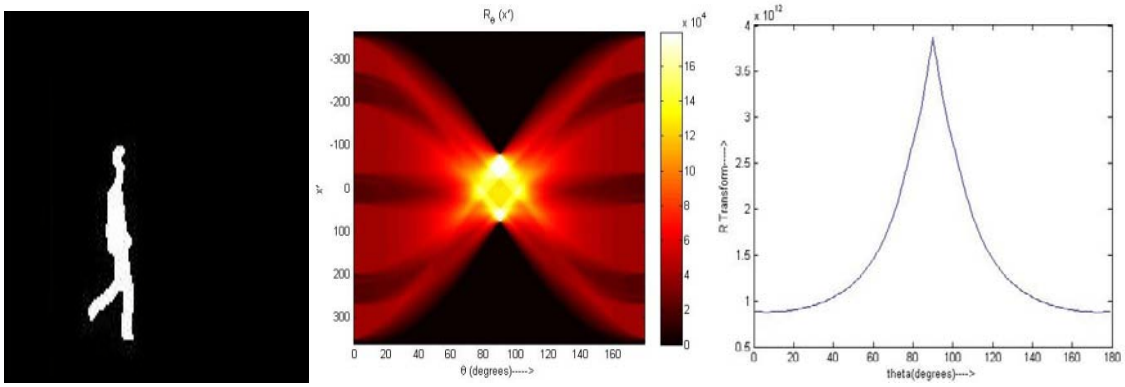
(a) Original Image



(b) Rotated by 30^{o} anticlockwise



(c) Scaled Image



(d) Translated Image

Fig 4.4: Shows 1D R transform geometric profiles of human running silhouette which has been rotated translated and scaled silhouettes. In the second image we find the more change in the brighter portion than the other images because we find more variation in values corresponding to projection lines of RT. The magnitude of translated images varies compared to scaled image and signal representation remains the same.

Normalization of R transform will improves the similarity measure and represents the features in a very compact manner. The normalized form of R Transform is defined as [55, 59].

$$R_{norm}(\theta) = \frac{\int_{-\infty}^{\infty} R(\theta)d\theta}{\max(R(\theta))} \quad (26)$$

4.3Methodology of Motion Features

Our method of computing the R-Transform is similar to the approach defined in [59].

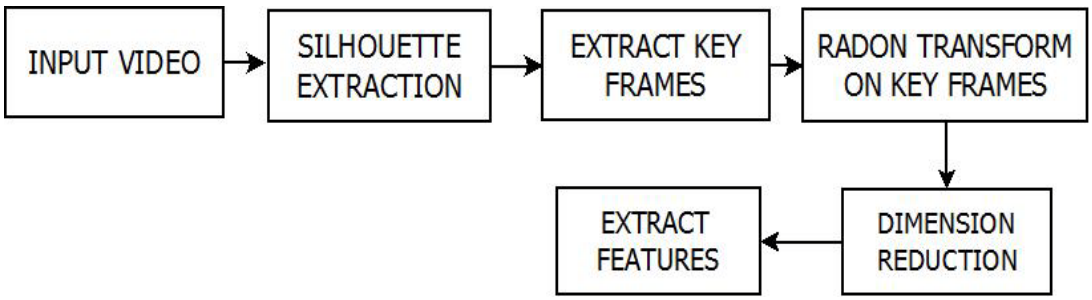
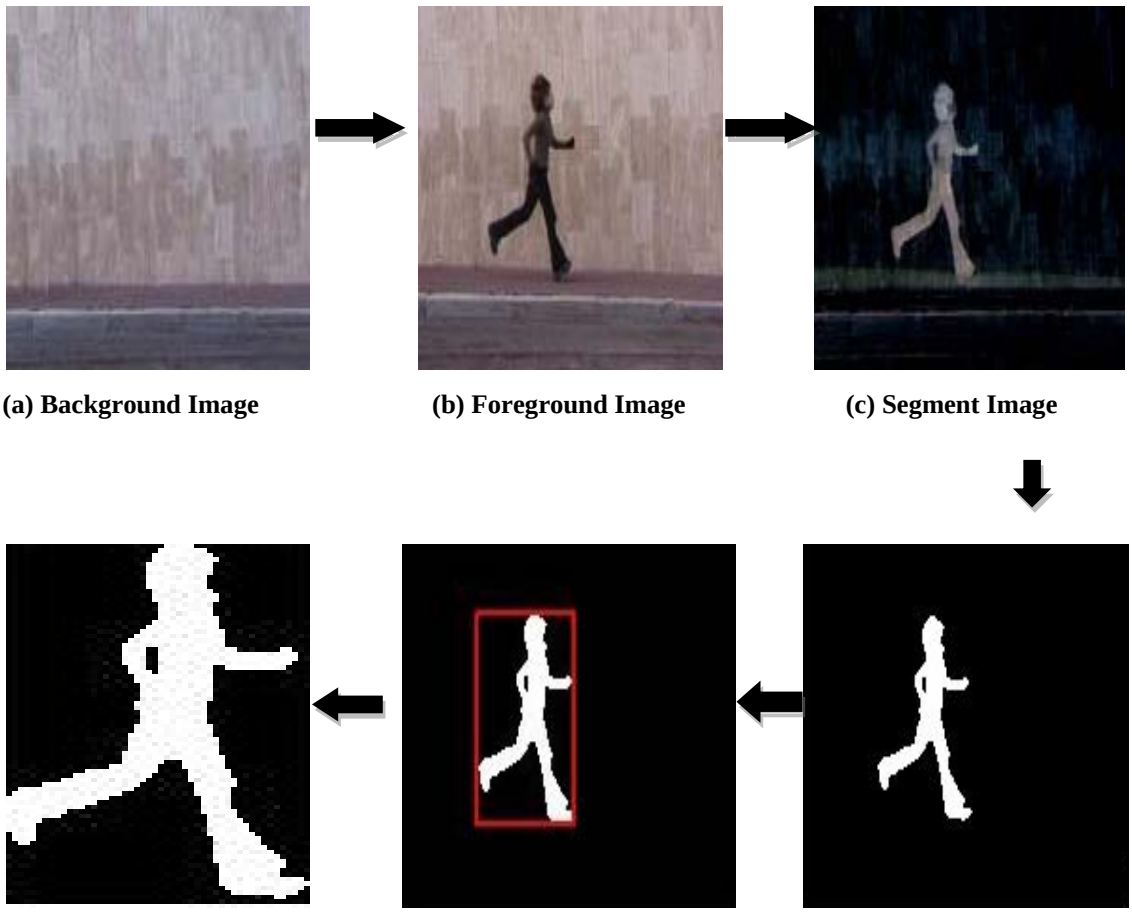


Fig4.5: Representation of extraction of motion features from Human action videos.



(a) Background Image **(b) Foreground Image** **(c) Segment Image**

(f) 50×50 Normalized ROI **(e) ROI Image** **(d) Silhoutte Image**

Fig 4.6: Procedure of finding a normalized ROI binary image

The radon transform applied to the ROI of binary silhouette images, but before employing it, we extract the key frames which represent far most discriminating features. We extract these frames from the sequence of frames by using Area of Binary image as discriminative feature. We first calculate the area of all the silhouette sequences and averaged to find the mean value.

$$Mean = \frac{[\,A_1, A_2, A_3 \ldots\ldots\ldots\ldots\ldots\ldots A_n\,]}{n} \quad (27)$$

WhereA_1, A_2, A_3...A_n represents the area of silhouettes obtained from video and 'n'defined as the number of silhouette. Then we extract the frames having area more than the mean value.

4.3.1 <u>Algorithm:</u>

Step 1: Take input video

Step 2:for i=1 to the number of frames.

Step 3: Extract the silhouette from Background subtraction technique.

Step 4: Apply median filter for smoothing.

Step 5: Convert frames to binary.

Step 6: Calculate the area of each silhouette.

Step 7: Compute the mean value

$$Mean = \frac{[A_1, A_2, A_3 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots A_n]}{n}$$

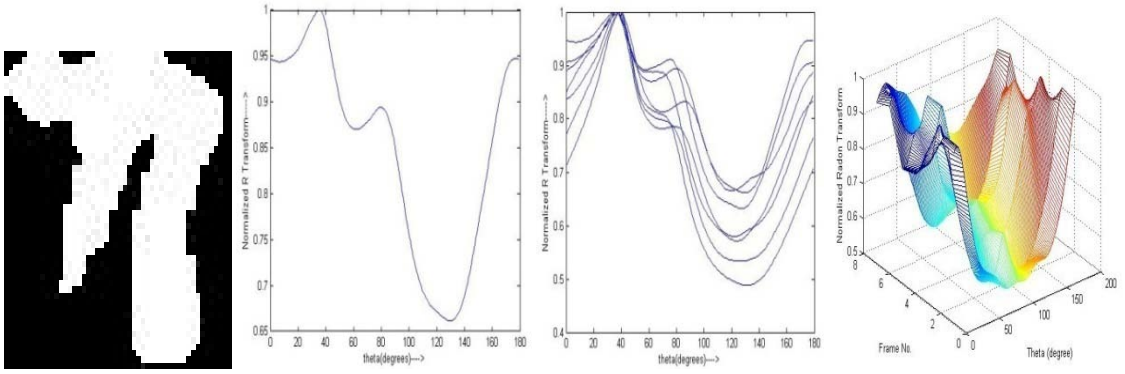
Step 8: If area of silhouette greater than mean value, then Key Frame.

Step 9: Otherwise EXIT

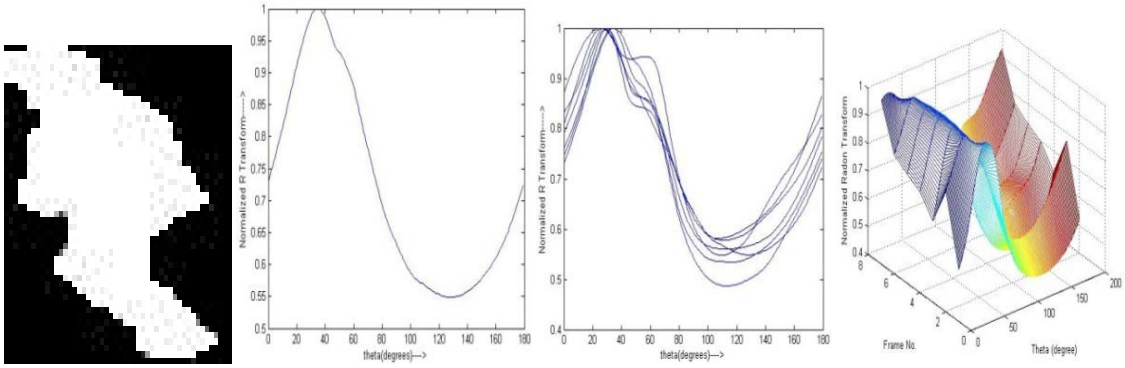
After applying the key frame algorithm we reduced the frames of video sequence to 7 frames. We apply the Radon transform on our extracted frames on all the 180 angle degrees and 2-D representation of feature vectors. Column vector represents the radon coefficients and a row vector defines the features. It will give directional feature vector matrix of 7 × 180 dimensions. This 2-D representation of radon coefficients converted into a 1-D representation of R-transform 1×180 feature vectors. Normalization will be

performed according to the equation (26). We then performed the dimension reduction techniques on 1×180 feature vectors so as to get the reduced discriminative feature vector of 1×7. We will study the various dimension reduction techniques in the next chapter. Here, we are showing R-transform signal for the some of the human action activities.

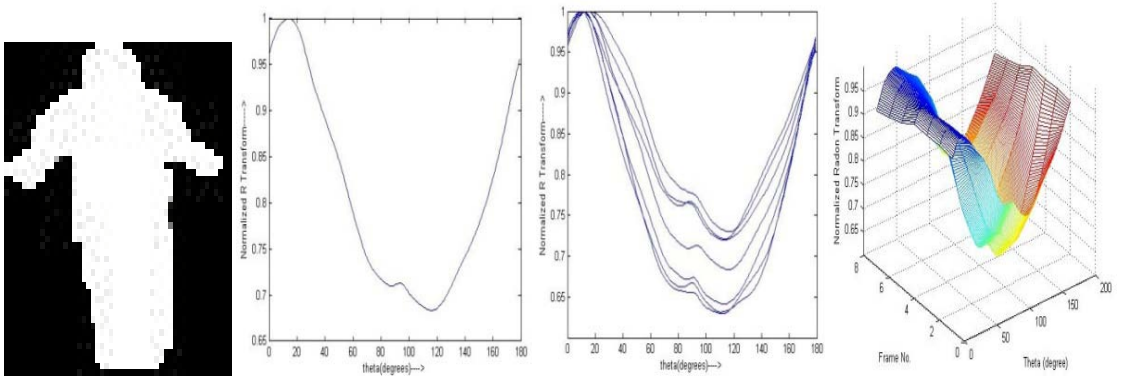
Representation of R-Transform Signals



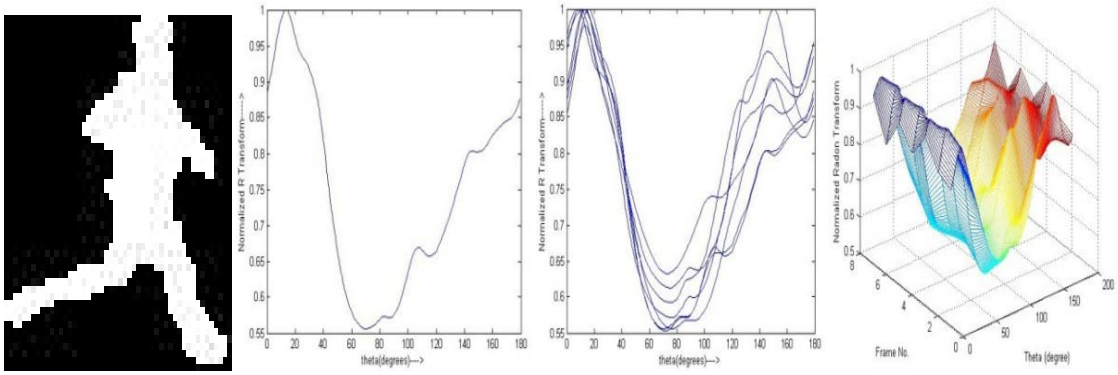
(a) Bending



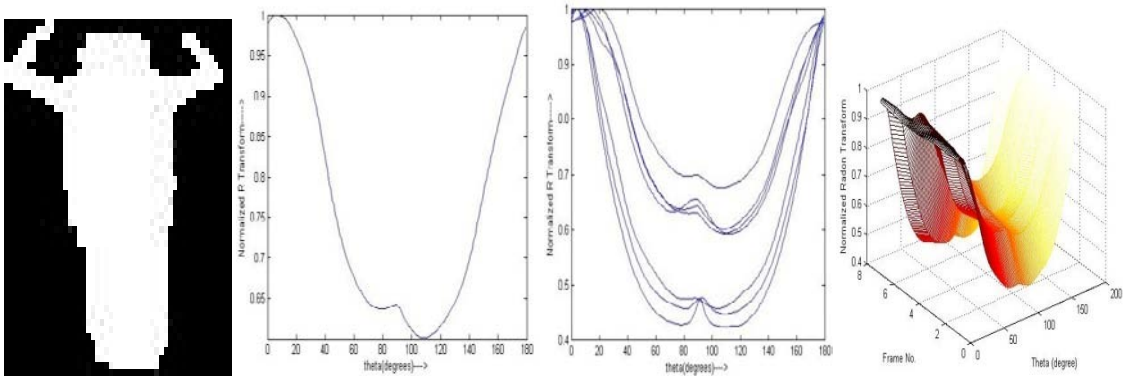
(b) Jump



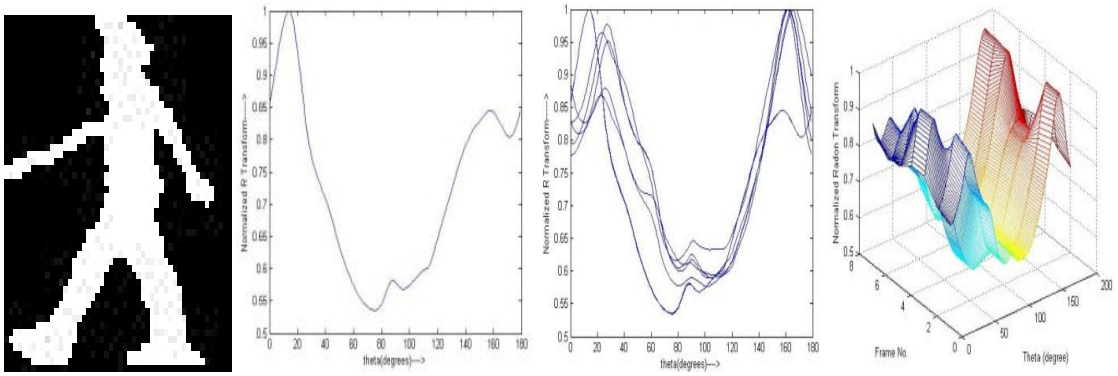
(c) Jump in Place



(d) Running



(e) Two Hand Wave



(f) Walking

Fig 4.7: There are four images in a row, first image represented as 50 ×50 normalized image. The second image shows the Normalized R- transform of 50 ×50 images. Third image shows the R-transform for multiple key frames represented by same class activity performing by different people. The x-axis represented by theta variation from 0 to 179 degrees and y-axis depicts the normalized R-transform. The fourth image represents the three dimension mesh plot of the R-transform signal. Numbers of orientations are represented by X-axis; whereas, the number of Key frames is represented by Y-axis and z-axis the normalized R- transform signal.

From the above figure it is clear that similar class activities have the same R-transform signal. Variations in the activity representation showed the corresponding change in R-transform signal. Therefore R-transform signal can be used as one of the discriminating motion representation signals.

4.4 Advantages of Radon Transform

Compared to the other feature representation techniques such contour based, Zernike moments, R- transform represent the shape of the information by calculating the pixel variations in different angles. Therefore, it gives us more local information about the object. It is insensitive to noise; scaling and translation invariance properties will further improve the effectiveness of this method

Chapter 5

Dimension Reduction

When we have a large set of features at that time we face problems in the recognition process as it takes a long time in classification and also increases the complexity of the system. Therefore, we prefer to reduce the dimensions of the features set to lower dimension subspace without any loss of information [60]. Consider we have obtained feature set $X = \{x_i \in R^d\}_{i=1}^N$ Where 'd' represents the higher dimensional space with or without the label set, we set this project into new feature vectors $Z = \{x_i \in R^p\}_{i=1}^N$ That not only has lower dimensional space 'p' ($p \leq d$), but also preserves discriminating characteristics of the original features. Dimension Reduction (DR) is widely used in the classification, reduced feature representation and regression models. In this chapter, we will study about the brief introduction regarding dimension reduction techniques that we used in our work, their algorithms, advantage & disadvantage. Experiments of DR methods are performed on the R-transform features that were explained in the previous section. Results of DR techniques are shown in the Experiments and Results section.

DR techniques are divided into two groups (1) **Linear dimensionality reduction (2) Non-linear dimensionality reduction.** These are as follows:

Linear Techniques

- Principal Component Analysis (PCA)
- Singular Value Decomposition (SVD)
- Factor Analysis (FA)
- Locality Preserving Projections (LPP)
- Independent Component Analysis

Non-linear Techniques

- Isomap
- Kernel PCA (KPCA)
- Laplacian Eigenmaps (LE)
- Local-Linear Embedding (LLE)

5.1 Principal Component Analysis (PCA)

PCA [61] which is also known as "Karhunen-Loeve Transform" is the unsupervised DR technique which reduces the dimension of data in linear subspace. Unsupervised means that it does not take any information regarding training labels (input data). The dominant features of dataset are obtained from PCA by solving the Eigen value problem of Covariance matrix (C) of the dataset, represented as

$$k = U^T CU, \text{ where columns of U, } u_i, \text{ are Eigen vectors.} \quad (28)$$

5.2 Linear Discriminant Analysis (LDA)

LDA is a supervised and efficient discriminating DR technique which preserves the relation between features and training labels by taking account the information of training labels (input data) [65]. Similar to PCA, LDA also generates the linear transformation matrix which here seeks projection that separates different classes as far as possible$W^T S_B W$ and maintains closeness within the class$W^T S_w W$. The desired projection transformation matrix given as $z = W^T x$

LDA is a supervised learning technique therefore it improves the accuracy in recognition because the system already knows about the features information. But LDA has certain limitations (1) LDA is dependent on the mean information therefore mean should have sufficient critical information for discrimination. (2) It is a linear method therefore it does not work for the non-linear distributions. (3) To improve the accuracy of LDA we require the large training set [62,63,64].

5.3 Kernel PCA (KPCA)

KPCA is the non-linear technique which uses the Kernel trick to represent the features in higher dimensional space**.** Then it extends the linear concept of PCA, which extracts the principal components of, variables that are non-linearly related to the input variable [60]. In kernel-based PCA initially we use the non-linear function $'\phi'$ in transforming the input vector $X = \{x^{(n)} \in R^d\}_{n=1}^N$ on to new vector ϕ (X) which has higher dimension than X after that we use the PCA to project the data into lower dimension space [64]. The kernel function mapping defined as

$$K(y, z) = \varphi(y)\varphi(z)^T \quad (29)$$

Kernel PCA helps in improving the linear discriminability but the size of kernel matrix also increases.

5.4 Local Linear Embedding (LLE)

LLE is defined as the unsupervised manifold learning based dimension reduction method. Unlike PCA, LDA, KPCA which is based on the maximization of the variance, LLE searches the nearest neighbours around the data point in high dimensional space. For each nearest neighbour weight are provided which describe the data point linearly with respect to neighbours. The weights are assumed to be invariant against translational, rotation and scaling parameters which keep the local properties same to the original space. These weights will reconstruct the data points into other 'D' dimensional space. From the 'D' dimensional space we mapped the features into lower dimension'd'$(d \leq D)$. While transforming the features data from one dimensional space to another it maintains the local characteristics of the features.

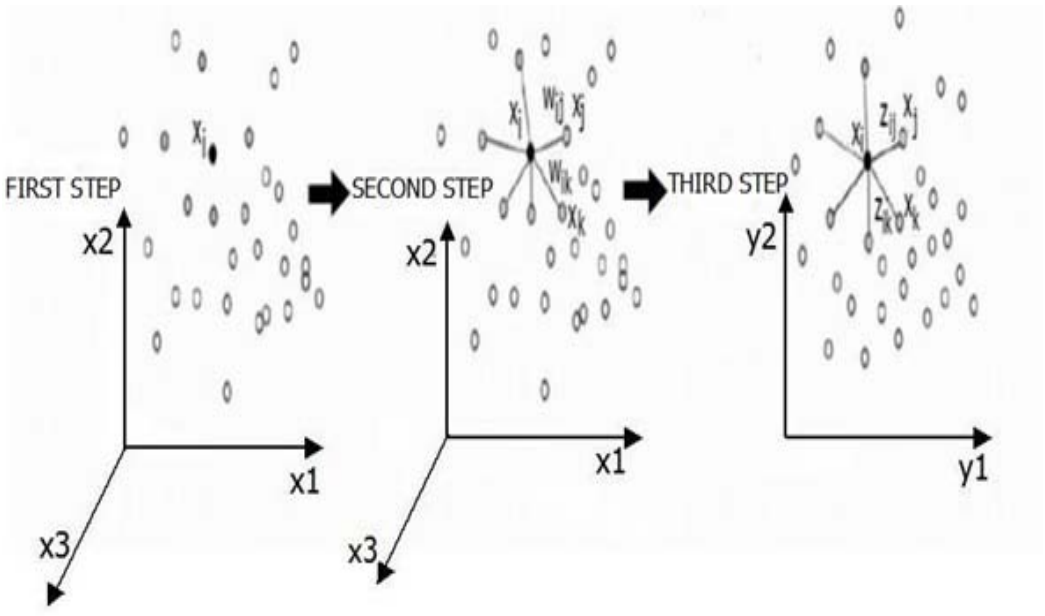


Fig 5.1: Representation of LLE algorithm procedure [65].

Mathematical Terms [63, 65, 67]

- Given a training set vector$X = \{x^{(n)} \in R^d\}_{n=1}^{N}$, of $d \times N$ dimension.

- Construct the weight matrix 'W' to record the linear reconstruction coefficients, where the i^{th} row $\left[w_{ij} = 0\right]_{1 \le j \le N}$ is the initial coefficients of sample the'i'from its neighbours.
- The set containing all the neighbours of $x^{(i)}$ is denoted as Nei(i)

$$N_i = |Nei(i)| = \text{neighbours of } x_i.$$

- Define linear coefficients $w^{(i)}$for each sample i with the least reconstruction error:

$$\varepsilon(w) = argmin_w \left\| x^{(i)} - \sum_j w_{ij} x_j \right\|^2 \quad (30)$$

- Two constraints are set on $w^{(i)}$
  - $w_j^{(i)} = 0, \quad if j \neq Nei(i)$ (Neighbourhood constraint). It means that $x^{(i)}$ is only reconstructed from its neighbour.
  - $\sum_{j=1}^{N} w_j^{(i)} = 1$ (For translation invariant).
- Re-embedding in the reduced feature space -After the previous step, now we get a matrix W which contains the information of local linear reconstruction coefficients. Now we also expect these coefficients are preserved in the reduced feature space Z. To achieve this, the following function is desired to be minimized:

$$\phi(Z) = \sum_{i=1}^{N} \left\| z^{(i)} - \sum_{j=1}^{N} w_{ij} z^{(j)} \right\|^2 \quad (31)$$

Step 1: Define the neighbour of each sample

Algorithm:

For 1: N

define a 0 ×1 $h^{(i)}$ vector to record the neighbour indices of $x^{(i)}$

define a d× 0 $\Gamma^{(i)}$ matrix to record the neighbour vectors of $x^{(i)}$

define a N_i ×1 $\eta^{(i)}$ vector to record the coefficients of $x^{(i)}$ coming from $\Gamma^{(i)}$.

for 1:N

If ($x^{(i)} \in Nei(i)$ and $i \neq j$)

$f^{(i)} \longleftarrow f^{(i)T} j]^T$

$\Gamma^{(i)} \longleftarrow [\Gamma^{(i)} x^j]$

End;

End;

End;

Step 2: Define the linear reconstruction coefficients

Algorithm:

for i= 1: N

$$C^{(i)} \leftarrow \left(x^{(i)}1^T - \Gamma^{(i)}\right)^T \left(x^{(i)}1^T - \Gamma^{(i)}\right)$$

$$\eta^{(i)} = \frac{C^{(i)-1}1}{1^T C^{(i)-1}1}$$

for j=1:N_i

$$w^{(i)}_{h_j^{(i)}} \leftarrow \eta_j^{(i)}$$

End

End

$$W = [w^1, w^2, \ldots\ldots w^N]^T$$

Step 3: Re-embedding in the data points in reduced feature space:

Algorithm:

(1) Define a N×N matrix M= $(I_{N\times N} - W)^T (I_{N\times N} - W) = (I_{N\times N} - W^T - W + W^T W)$

(2) Perform EVD on M , MV=VΛ, $M = V\Lambda V^T$ *, where Λ is in descending order.*

(3)$Z = \left[v^{(N-p)}, v^{(N-p+1)}, \ldots\ldots v^{(N-1)}\right]^T = V\left[O_{p\times(N-1-p)} \middle| I_{p\times p} \middle| O_{p\times 1}\right]^T$,

where $O_{p\times(N-1-p)}$, $meansap \times (N-1-p) zeromatrix$

This method is better in representation of non-linear distribution data points and also maintains the local structures, but it is computationally expensive [67]. LLE is dependent upon the representation of data points in higher dimensional space; if they are different in representation then LLE can't ensure their reconstruction in lower dimensional space.

In our method we reduce the dimension of R-Transform by LLE dimension reduction technique.

Chapter 6

6.1Action Recognition from Still Images

Recently in human action recognition, still-images have shown great scope in computer vision and pattern recognition. It focuses on identifying a person's action or behaviour from a single image. Compared to the spatio-temporal techniques, still images do not require any temporal information therefore working on the still images is easy. Wang et al. [68] introduced the concept of action recognition based on still images. They used the canny edge detector to represent the human action image and features are clustered into similar body poses. To compute the clusters, Spectral clustering method is used which compares the pair wise images. Li et al. [69] proposed the concept of "exemplarlet" which contains enough visual information to identify the human action. These "exemplarlet" are manually selected and segmented used for training of data. Latent Multiple kernel learning method is employed for the learning the action classifiers. This method does not work accurately for complex images. Yao et al. [70] used the concept of random forest decision tree algorithm to search for the discriminating patches of the human action region. Yao et al. [71] used the characteristics models and parts of actions to describe the poselet of the actions. Li and Fei-Fei [72] introduced the integrated method which is based on the appearance information on image and occurrence of action scenes. They segment the still image information into smaller subparts for e.g. an image representing a rowing scene, it will observe into subparts such as athletes' actions, boat and river information. Thurau and Hlavac [73] proposed the method based on pose information of human action. They first extracted the region of interest (ROI) from still image or image sequence which represents the pose. Histogram of gradients with non- matrix factorization (NMF) representations is computed to represent feature vectors. Classification of actions based on by means of histogram comparison. Anna Lopes and Santos [74] proposed the transfer learning approach where contextual information from still images applied to the test video sequence. Video sequence represented by the set of frames (summary frames) and information from still images compared to these set of frames. Zheng et al. [75] introduced the similar approach [74] combination of poselet and contextual

information for recognizing human action from still images. HOG descriptors are calculated from the pose representation and contextual information represents the occurrence of the action in an image. Sparse coding is used to form the visual vocabulary instead of k-means as it provides a compact representation of features.

Hui et al. [76] introduced the spatial pose based exemplars to represent the human object interaction (HOI) from still images. Exemplars are the probability density function that gives us the idea how the object with respect to atomic pose. Atomic poses referred to the set of images/poses that often occur in (HOI) activities. Exemplars are computed on each pair of object and atomic poses. This method represents the mutual spatial structure between the object and human. Sharma et al. [77] proposed the new model where they recognize the human action on the basis of "Expanded parts model". In this model, object is represented by collection of part templates which provides discriminating description of the appearance of objects. Recognition of an image is based on scoring function which is observed from learnt part templates. This method gives more discriminating parts of object without including background which improves the accuracy of the system. Liu and Shao [39] presented Adaboost learning algorithm in the selection of a set of the most discriminating key frames described by the Pyramidal Motion features (PMF). They used the optical flow algorithm to extract the motion features initially and then form the intensity images at different scales. Final motion vector comprises PMF of two adjacent frames. Chaaraoui et al [78] used the key poses concept for action recognition. Key poses are extracted from the sequence of frames by using k-means clustering algorithm with the Euclidian distance. Dynamic Time Warping (DTW) matching algorithm is used to classify the test sample from training set images. As this approach is not temporal based therefore it is simple and noise independent. Its main drawback is that although it gives an idea about action, i.e. whether it is walking or running but does not give any information about the motion e.g. walking forward or backwards. Understanding the human action from still images has always been a challenging task. Still images do not provide sufficient information of human action compared to the motion information, where as a sequence of images or videos describes this information. Moreover, in case of still image we have to estimate the position and posture of the object. Therefore, we try to recognize the human actions, including the motion information with still images. Our approach starts with finding out the still images from videos and then employing the pose extractor

representing the pose via pyramid of histogram gradients. Motion information is extracted from the silhouettes by calculating the Radon coefficients. This approach will give more global information of the object as it covered both the local information of the posture and global information about the motion. For the classification we used the Support vector Machine (SVM) and K-Nearest Neighbour (KNN) algorithm.

6.2 Still Image Extraction From Video Sequence

To represent the still images from the videos we have to extract these from the video sequence. These postures will clearly represent the actions performed in the video. There are many approaches to extract the key postures or discriminating images from video. First, we give an idea about the defined basic techniques that are earlier performed and then we represent our idea of extracting the still images from video sequence.

Some previous approaches

We can describe the key frame extraction algorithms into four groups: (1) Shot Segmentation (2) Clustering (3) Motion (4) Video Summarization.

1. **Shot Segmentation** - In this approach, video sequence is divided into shots based on the sudden change in the video. From each segmented shot category, we extract the key frame independently. Two approaches can be used for the selection of key frames. First, from each segment group key frame is represented by the first frame of each group, this approach is simple, but keeping the first frame as key frame of each shot category is not a stable method and does not capture the discriminating feature of video content [79]. The second method is based on colour difference. In this method, we measure the colour difference between the frames, if there is variation in the similarity and changes are significant then consider that frame as key frame

2. **Clustering method** - In this approach similar group is arranged in one manner and from that group we choose the key frame. One method is frame by frame difference and if the there is a change while comparing then we called it as different category frames otherwise they are placed in the same category. Another method is forming cluster groups from video frames. From each cluster group we select the key frame. This method is efficient from the earlier method as it does not require temporal ordering of frames.

3. **Motion based approach**- In this method we first compute the optical flow frame wise frame of shot category and finds the motion metric function [80]. A key frame is represented by the local minima of metric function which varies with the time.

4. **Video summarization**- In this method, first video frames is sampled either randomly or uniformly. Then for this sampling distribution, we select the key frames. This method is simple and fast for a selection of key frames, but due to the sampling distribution sometimes distinctive frames are lost.

In this work, we proposed simple and effective approach for the selection of key frames from the action video sequences. There are two steps in the approach. Firstly, key postures are selected globally and then locally, we extract the most discriminated image (final key posture) from the globally extracted frames. For the first step, we first form the groups of frames as in the video; the action does not vary instantaneously. Then from each group we select the frames and convert into CIELab colour space. In this model 'L' stands for the luminance and 'a', 'b' components define the colour opponent dimensions i.e. simultaneously no two colour's appear at the same time. Compared to the other colour model (RGB, CMYK),CIELab colour space closely conforms to human perception of colours [81] due the 'L' component which has the same luminance as human's ability to see the things in the lightness.

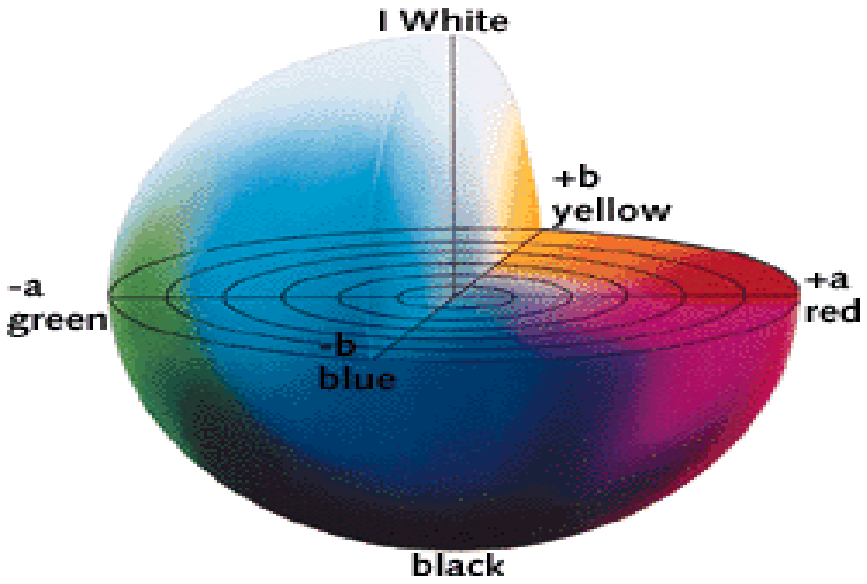


Fig 6.1: The CIELAB colour space

First, divide the action sequence into segments, then choose the first frame from each group and convert it into CIELab colour space. Compute the sum of pixel distances for all the three components 'L', 'a' and 'b' which is defined as

$$D_t = \left|\sum_{i=1}^{M}\sum_{j=1}^{N} S_{ij}^{t} - S_{ij}^{t+1}\right| \quad (32)$$

'S' is defined as the first frame of each group, 'M' and 'N' defined as rows and columns. Histogram difference typically considers global changes in a video sequence and is helpful to detect abrupt changes in the shot. On the other hand, a gradual change can be characterized by both global (fading) and local changes (wipe). Pixel difference is used to identify local changes in a sequence. For the selection of key frames from the groups we use the Fuzzy inference method. The distance that we computed for the subsequent frames is used to make the Fuzzy inference model. Fuzzy inference model we will study in the next section and will also discuss the fuzzy rules and its algorithm. In the next step, from the set of key frames we select the single frame which has a maximum distance compared to others.

6.2.1ONE Key Frame Selection Rule

When we apply the fuzzy rule, then pixel difference values are spread over the membership function. Frames are extracted by applying the fuzzy rules over the function. These sets of frames are then compared internally and ranked according to the difference values. Higher difference values show that corresponding frame has a higher variation or changes with respect to other frames. If we representf_1, f_2.....,f_n frames extracted by applying the fuzzy rule, Then single frame defined as

$$\text{Single Frame}=\text{argmax}\left(\sum_{i=1}^{n} f_i\right) \quad (33)$$

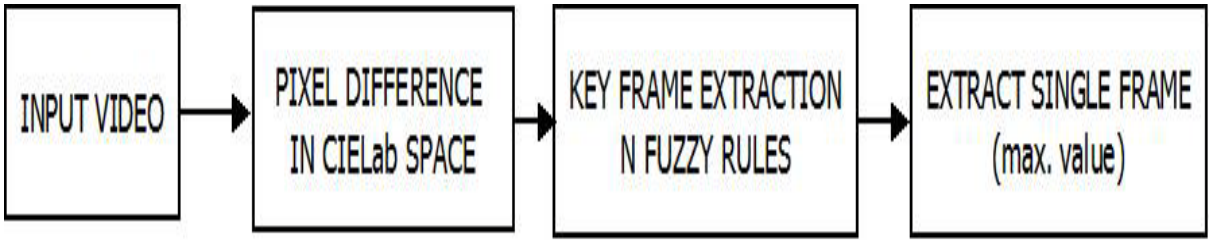


Fig 6.2: Flow method of Single Frame extraction

Fig 6.3: Frames extracted on the basis of fuzzy rules of the jumping jack activity Weizmann Dataset

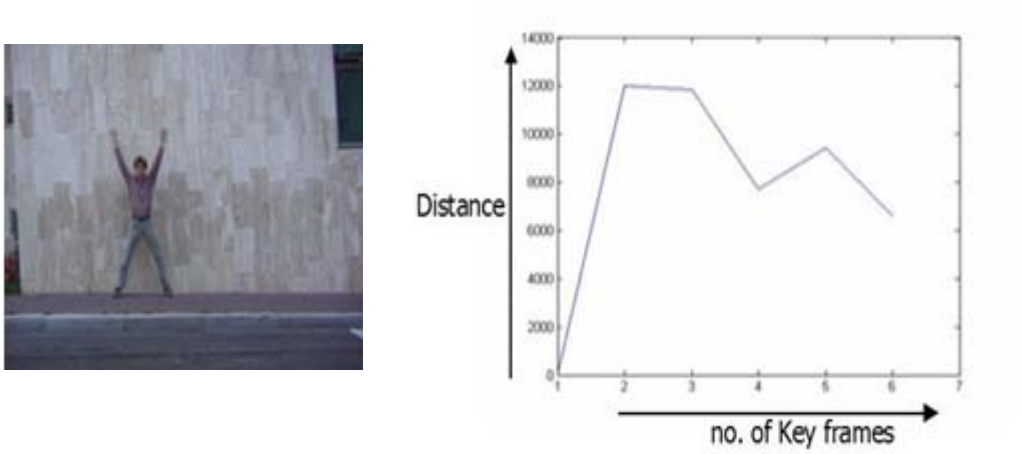


Fig 6.4: (a) Showing max difference, variation based key frame (b) plot of variation of difference with the extracted key frames

6.3 Fuzzy Logic System

The fuzzy logic system is based on the probability model where we have partial knowledge about the system rather than the accurate solution. It is derived from the fuzzy set theory, given by Lofti Zadeh in 1965 which is an extension of Boolean logic. Compared to the crisp logic, fuzzy logic always provides the measure of uncertainty. Fuzzy logic deals with the membership function which defines the probability of a variable belonging to a particular group. A membership function assigns a value between the 0 and 1. 0 membership value state that variable does not belong to the group and 1 explains the complete relationship to that group. As the value increases from 0 to1 then the degree of membership value also starts increasing for the variable. Consider an example of crisp function represented by Set A, this function assigns a value $\mu_A(x)$ to every $x \in X$such that

$$\mu_A(x) = \begin{cases} 1 & if and only if\ x \epsilon A \\ 0 & if\ and\ only\ if\ x \notin A \end{cases} \quad (34)$$

Crisp logic only gives hard decision whether it belongs to a group or not. In some of the video content of information may be large, so we cannot extract the key frames on the basis of distance metric. In this case some of the information regarding the frames may lose. Therefore, we use the concept of fuzzy system that will estimate the content of videos more globally. It is more flexible and it maps the set of inputs to outputs by using simple "if... Then" statements. The mathematical concepts behind the fuzzy reasoning are very simple. The performance of the fuzzy rules depends upon the choice of the fuzzy partitions.

6.3.1 The Design of Fuzzy Membership Functions and Rules

Membership functions are the deciding parameter of the Fuzzy inference system. There are different shapes of fuzzy membership function such as Triangular, Trapezoidal, and Gaussian etc. Other than choosing the shape of membership functions, setting the interval and number of functions is also an important parameter. For example, to model a temperature control system by Fuzzy logic, it is really important to know how many membership functions are needed (low, medium, high) and also choosing the intervals. These two factors also have a great impact on the outcome of Fuzzy logic system. Fuzzy rules can we write into a simple "if...else" statements. From these statements Fuzzy logic makes decisions about classifying the objects. We have observed and defined some fuzzy rules, but if type of video is not known, then rules are selected based on whether it is single shot or it contains multiple shots. From this set of fuzzy rules, we obtain outputs, which are used to make decisions.

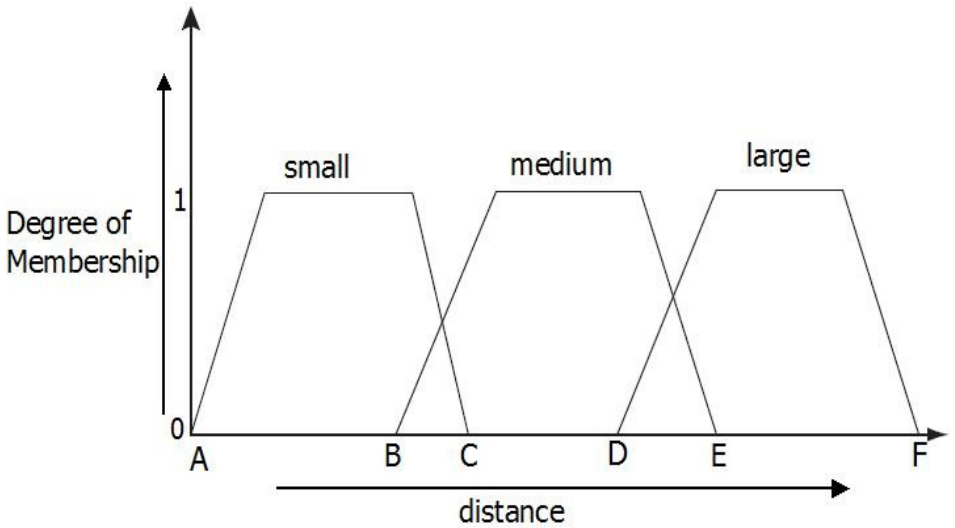


Fig 6.5: Fuzzy Trapezoidal Membership function

- **Rule 1 (Video type: sports video):** IF the distance between a segment frame and its neighbouring segment frame is “medium” THEN it is a key frame
- **Rule 2 (Video type: TV sitcom videos/movie video):** IF the distance between a segment frame and its neighbouring segment frame is “large” THEN it is a key frame
- **Rule 3 (Video type: TV sitcom videos/movie video):** IF the distance between a segment frame and its neighbouring segment frame is “small” THEN it is NOT a key frame.
- **Rule 4 (Video type: not known):** IF the distance between a segment frame and its neighbouring segment frame is “large” AND the video is of multiple shots THEN it is a key frame.

6.3.1.1 Algorithm:

Step 1: Input a video.

Step 2: Skip to every tenth frame.

Step 3: FFS is the first frame of each segment.

Step 4: Each frame is in CIELab colour space.

Step 5: Calculate the histogram distance for ‘L’, ‘a’, ‘b’ components

$$D_t = \left|\sum_{i=1}^{M}\sum_{j=1}^{N} FFS_{ij}^{t} - FFS_{ij}^{t+1}\right| \quad (35)$$

Step 6: Compute mean, ‘μ_d’ for all consecutive segment frame differences.

Step 7: This mean is then used to create membership functions dynamically.

Step 8: First medium is created by spreading it around the mean ‘μd’, and then small and large are created on either side of medium.

Step 9: Since values are spread over a range we used trapezoidal function

Step 10: The membership values are calculated as follows:

- $A = (\mu_d - \mu_d * 0.4), B = (\mu_d - \mu_d * 0.3), C = (\mu_d - \mu_d * 0.2),$ $D = (\mu_d + \mu_d * 0.4, E = (\mu_d + \mu_d * 0.5), F = (\mu_d + \mu_d * 0.8)$

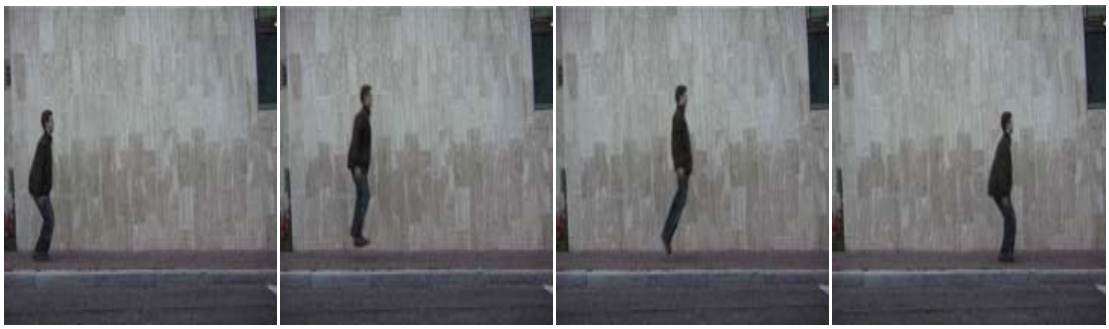
Working directly on still images offers great advantages as it does not involve the segmentation process, alignment of the images. We can directly apply our feature

extraction problem on these images therefore it reduces the time computation and complexity of the system.

Some of the examples of key frames extracted from the action videos:

(a) Jump_Jack

(b) Bending



(c) Jump

(d) Walking

Fig 6.6: Key Frames extracted on the basis of pixel differencing of Weizmann Dataset lying under Fuzzy membership values

Some of the examples of Single key frame extracted from the set of Key Frames:

Fig 6.7: Single Frame Extracted from the Key Frames of bending, jumping jumping_jack, walking, two hands waving, one hand waving and running.

6.4 Methodology in Still Images

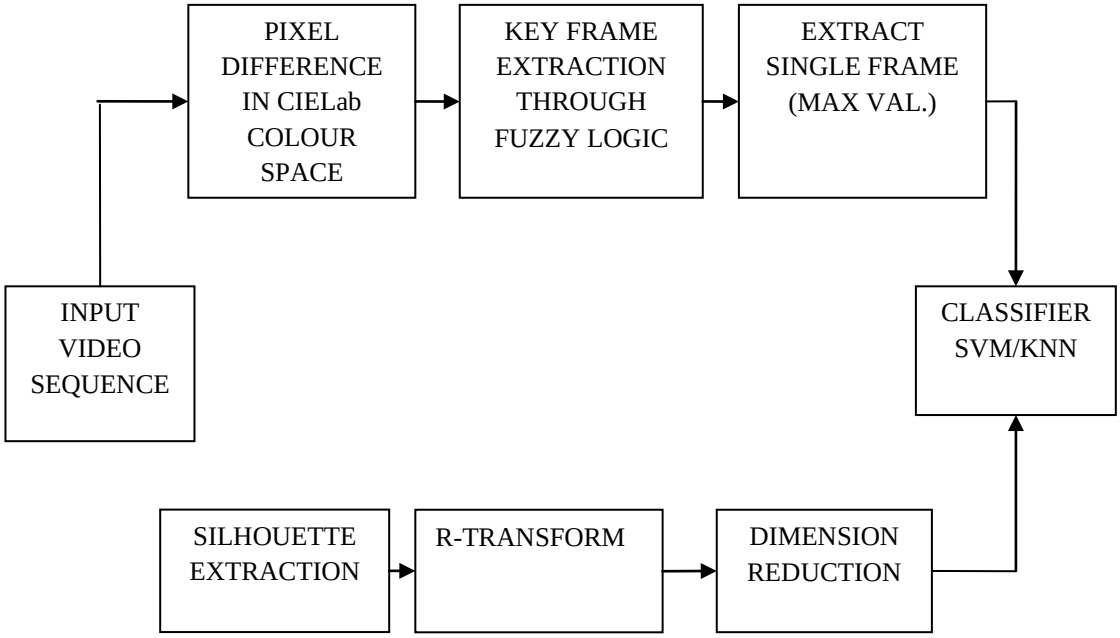
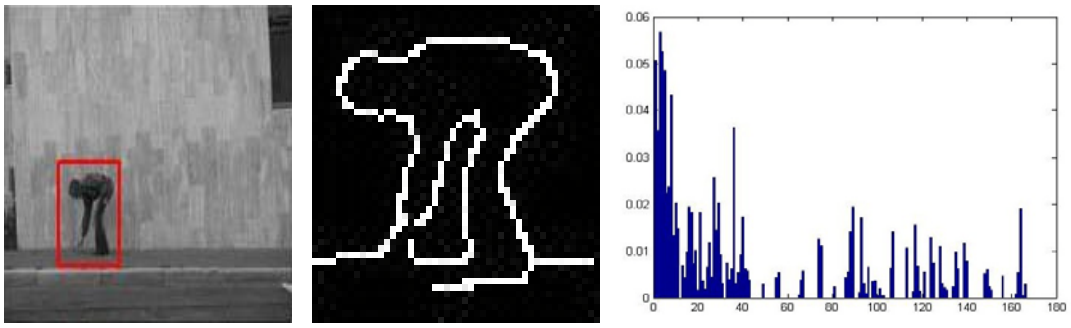


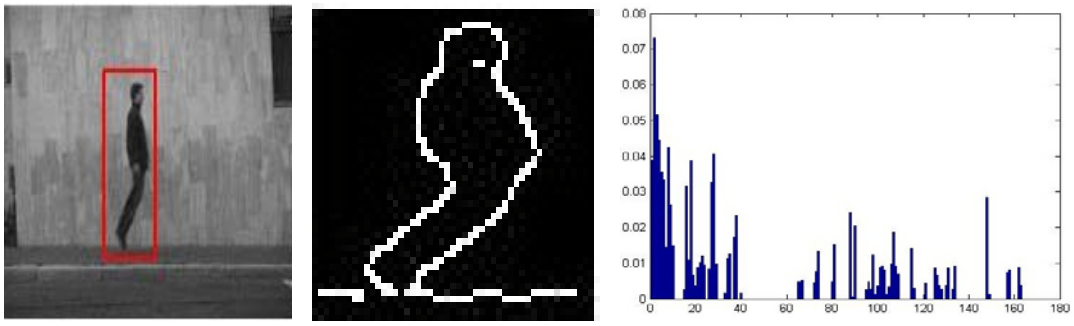
Fig 6.8: Representation of Spatial and Temporal approach

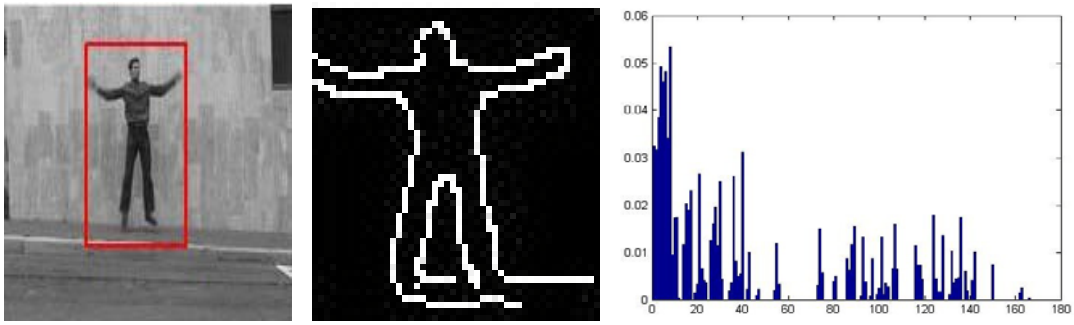
After extracting the single frame we extract the features from the posture representation. For feature representation, we use PHOG[45] which is already defined in the feature extraction method in detail. Motion information obtained from the Radon coefficients which are computed on the silhouette of the videos. This method provides combined local and global information of the system.

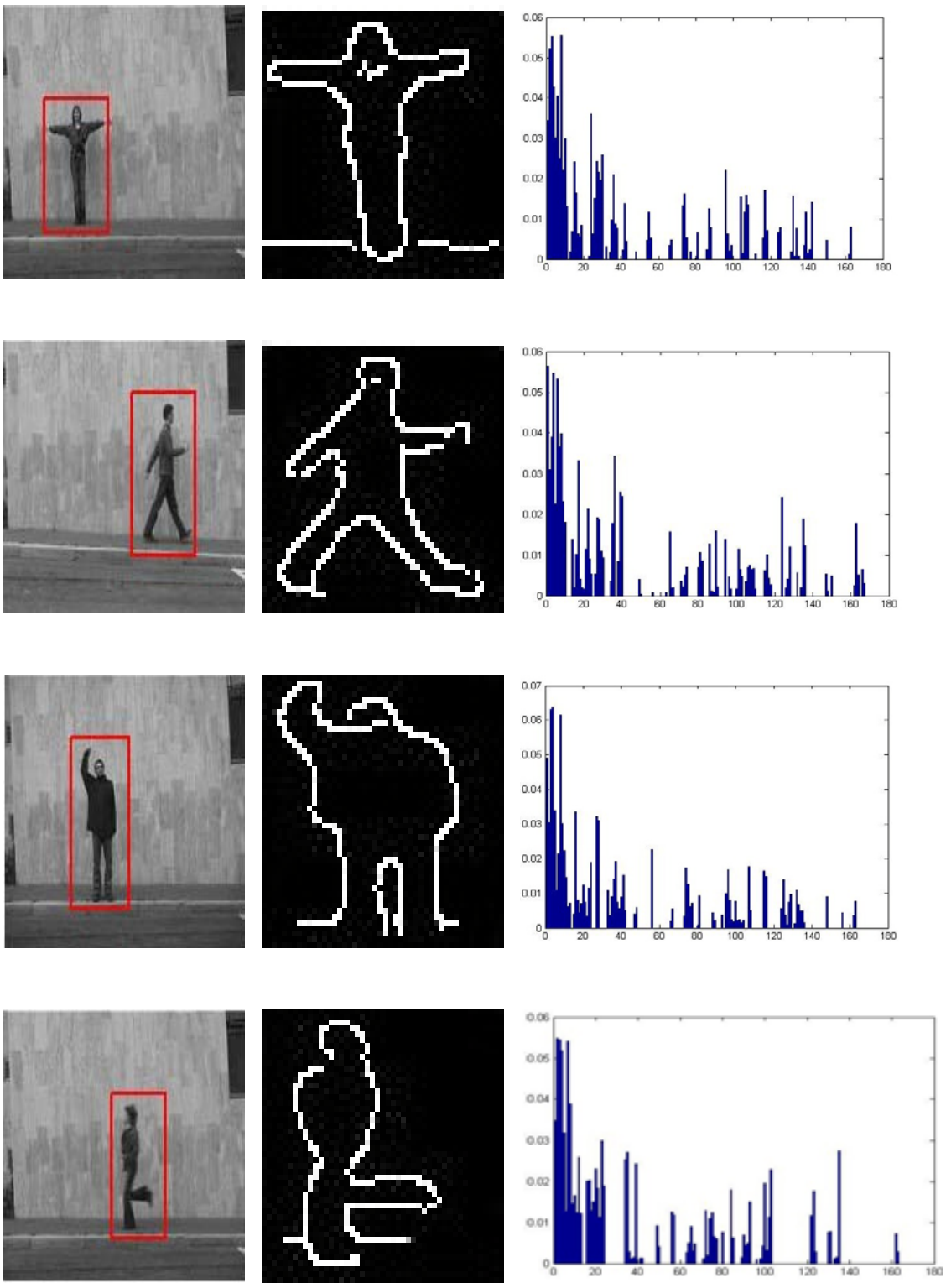
6.5 PHOG Apply On Single Frame

Pyramid of histogram gradients is a local approach it gives the information about the posture with more accuracy and is robust in nature. First, from an image, ROI is selected that will help to calculate the gradient on action not including the background. This is a simple and fast approach to calculate the features. For each action we get the separate PHOG vectors and characteristics of shape which is helpful in the recognition process.









(a) **(b)** **(c)**

Fig 6.9: (a) ROI Image of key frames, (b) canny edge detector of an image while computing PHOG, (c) Final PHOG descriptor. ROI helps in discarding the background variations and localized region around the object

Chapter 7

CLASSIFICATION

Classification is the process where objects are categorized into similar groups. There are two types of classification method in pattern recognition: supervised and unsupervised classification. In case of supervised learning classifiers learns from the input training data and then assigns the class label. The Unsupervised classification method does not learn from the input data, it finds the hidden layers in the unlabeled data by using clustering techniques. Supervised techniques, is mainly used for the object classification and detection while unsupervised is used for detection and segmentation process. In case of human action recognition, we have training data from which the classifier system learns about the training features and measure the similarity with testing features. There are many classifiers defined which help in the recognition of actions like Naive Bayes, k-nearest neighbor (KNN), Support vector machine (SVM) classifiers. In our work we recognize the human action by using KNN and SVM Multi-class classifier with leave out one method.

7.1 K-Nearest Neighbours

The k-nearest neighbour (KNN) is a simple machine learning algorithm for classifying objects based on a similarity measure (distance function). It is called non-parametric method as it does not learn an explicit mapping from training data. It simply uses the training data at test time to make the predictions. There are many distance functions which are used to measure the similarity, but it depends upon the type of features of the data. If the distance is small then there is more similarity between the two samples. Euclidean distance, Mahalanobis distance, Hamming distance are some of the common distance function. If the features are real valued then we use Euclidean and if the features are Binary valued then we use Hamming distance.

The Euclidean distance between any two points w = (x_1, x_2… x_{1n}) and z = (y_1, y_2… y_m) given as;

$$\mathrm{dist}(w, z) = \sqrt{\sum_{i=1}^{n=m} (x_i - y_i)^2} \tag{36}$$

The basic steps of the KNN algorithm are:-

- Find out the distances between the test samples from each training data.
- Sort the distances in increasing order.
- Choose the value of k having least distance value.
- For the classification of data voting principle is used.

Consider an example of two class variables: Class A and Class B represented by '×' and 'O', respectively. We want to classify a test sample, represented by '+', as whether it belongs to Class A or to Class B. In this case test sample will belong to that class A because its neighboured distance is least among k-nearest neighbours.

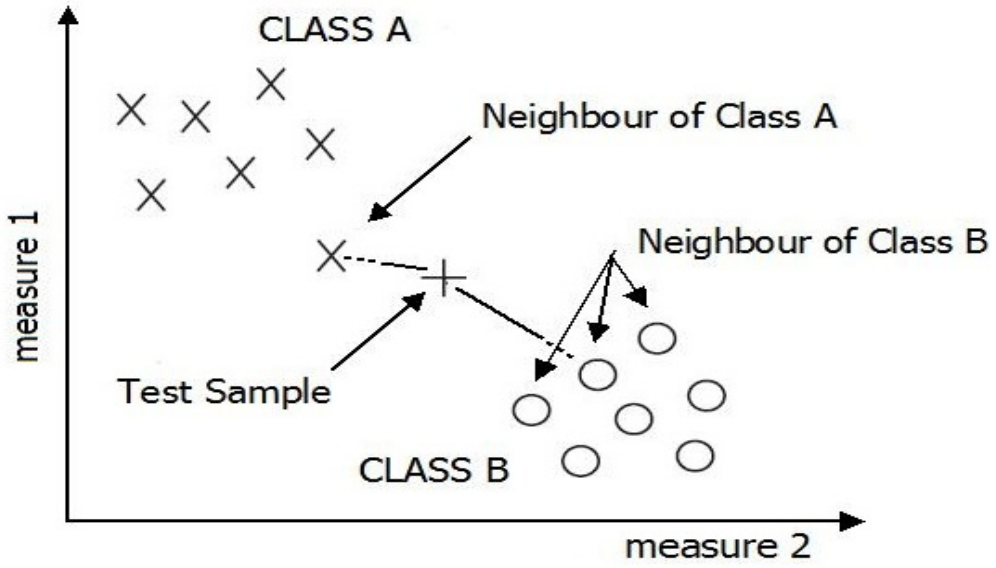


Fig 7.1: Representation of k-nearest neighbour clearly shows that test sample belongs to Class A as its neighbourhood distance is less compared to Class B.

KNN classifier is simple and easy to implement. It gives satisfactory results in the case of large training data with high value of 'k' and it uses the local information of data which makes it highly adaptable. But, in case if training set is large, KNN requires large memory and its prediction accuracy ability quickly degrades. Due to large amount of training data its computational cost is high.

7.2Support Vector Machine

SVM algorithm is based on the supervised learning [40] i.e. it has prior knowledge about the features which is to be used to predict the state of the class labels. It

transforms the set of vectors into higher dimensional space where it continuously finds the optimal hyperplane that separates the class labels. The hyperplane is called as a decision boundary which distinguishes the two classes and maximizes the separation, margin between itself and those lying nearest to it. These nearest set of points called as the Support vectors.

Mathematical Terms

- Given a training set vector$X = \{x^{(n)} \in R^d\}_{n=1}^{N}$ with class labels Yi ε (-1, 1).
- The optimal hyperplane follows the equation

$$y_i(w^T x_i + b) > 1 \; \forall i \tag{37}$$

- The optimal hyperplane is developed by solving the optimization problem;

$$\min \varphi(w) = \frac{1}{2}(w \cdot w)$$

- b is defined by the Karush Kuhn Tucker condition

$$L(\alpha) = \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i^T x_j \tag{38}$$

in related to $\sum_i \alpha_i y_i = 0 \; \alpha_i \geq 0 \; \forall i$

if$\alpha_i$is the solution of eq. (38) than optimal hyperplane is defined by [40]

$$w = \sum_i \alpha_i y_i x_i \tag{39}$$

The margin of x_i with respect to linear classifier is defined as the perpendicular distance x_i from the hyperplane w ⋆ x +b=0; in above figure it is defined by ‘n’.

Margin width defined as the margin between the two classes, measured perpendicular to the hyperplane.

$$M = (x^+ - x^-) \cdot n$$

$$M = \left((x^+ - x^-) \cdot \frac{w}{\|w\|}\right) \quad = \frac{2}{\|w\|} \tag{40}$$

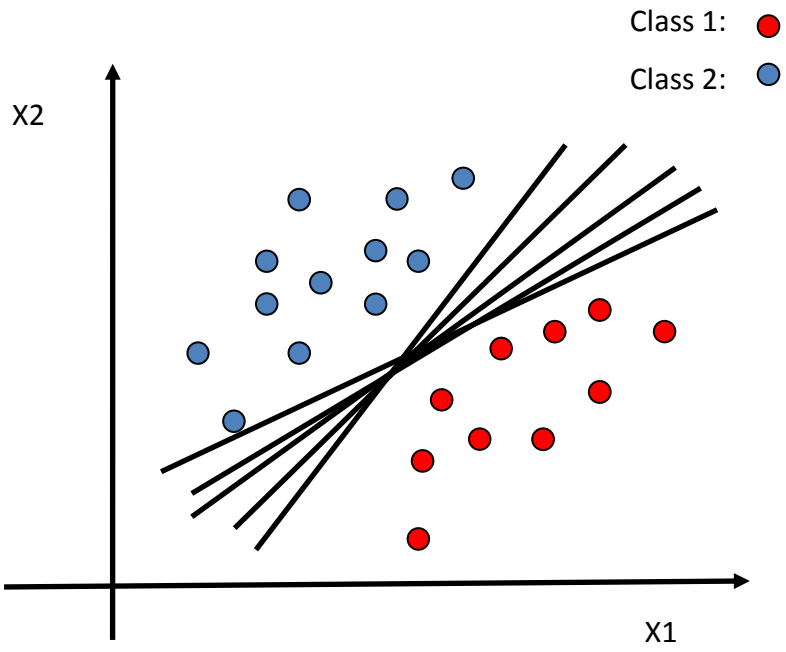


Fig 7.2: Several possible hyperplane for class separation

In case of non-perfectly separable, misclassification error of training sample has to be minimized. The minimal optimum solution expressed as

$$\min \varphi(w) = \frac{1}{2}(w \cdot w) + C \sum_i \xi_i \tag{41}$$

related to

$$y_i \left(\left(w \cdot \emptyset(x_i) \right) + b \right) \geq 1 - \xi_i, \quad \forall i$$

C is defined as the penalty parameter which controls the generalization ability of SVM and ξ is non-negative slack variable. The higher value of C results the over fitting problem on the training sample.

In case of non-linear SVM, a mapping function '$\emptyset$'is used to transform the input feature vector into higher dimensional feature space and further hyperplane distinguish feature vectors in separate classes [82].

$$x \rightarrow \emptyset(x), \quad \emptyset(x) \text{ is the mapping function.} \tag{42}$$

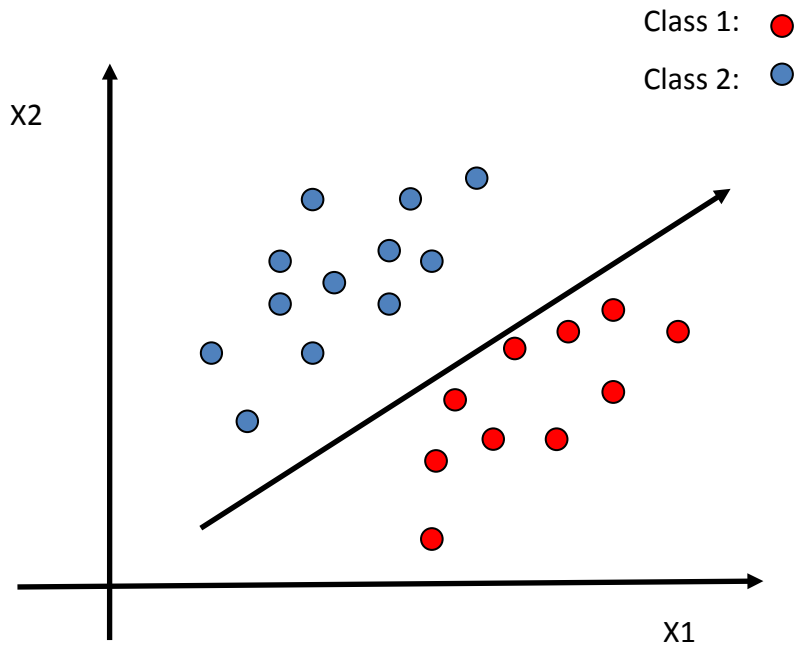


Fig 7.3: Plane is separating the Two classes

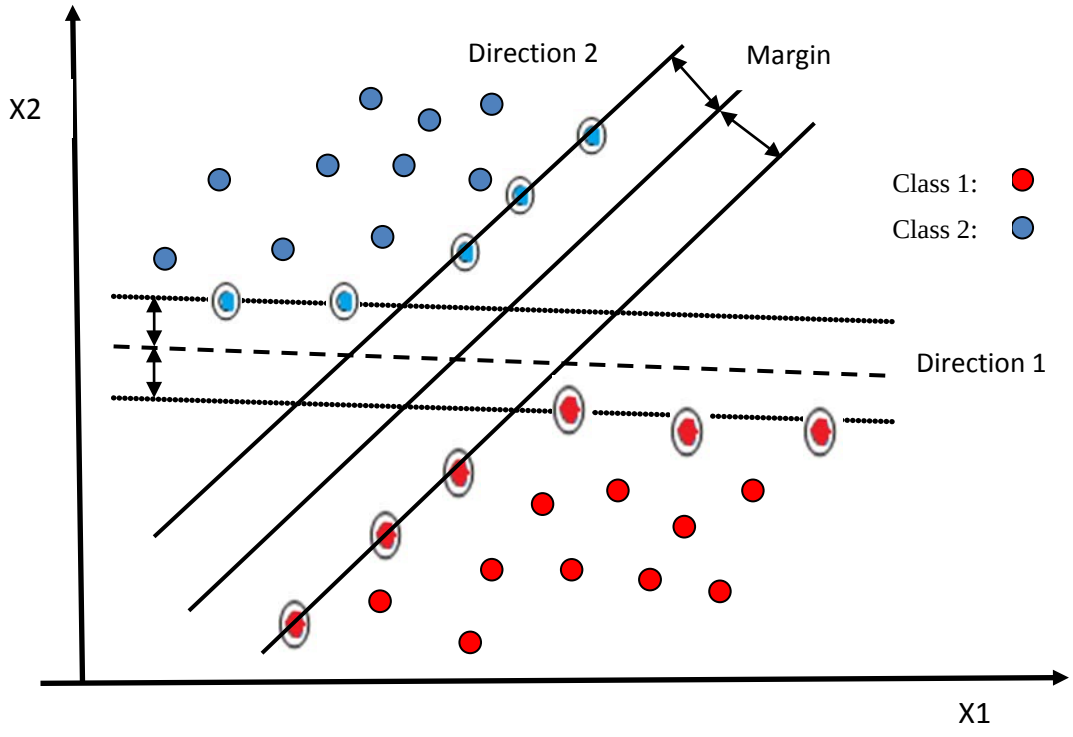
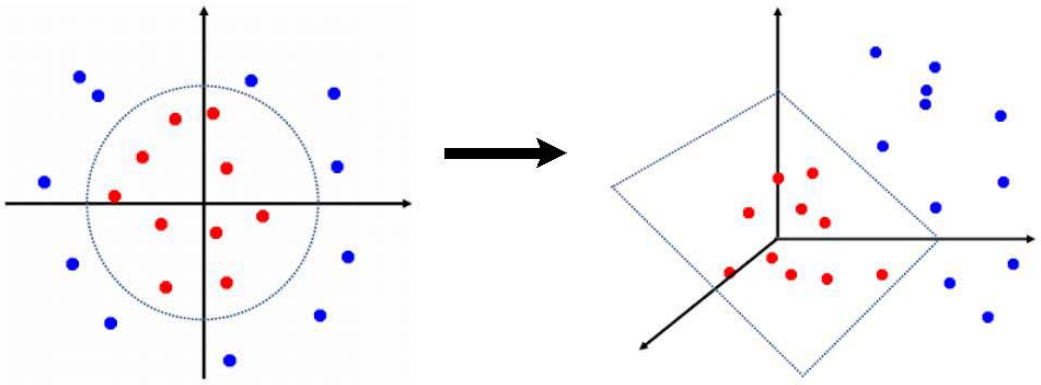


Fig 7.4: SVM classification example, where hyperplane separates the two classes. Points lying near to hyperplane called as the Support vectors.

This mapping function also called as kernel function. For eg: - Radial basis function (RBF), linear kernel and polynomial kernel.



(a) (b)

Fig 7.5: (a) Non linear representation of feature vector (b) mapping function transforms into higher dimensional space. Hyperplane is constructed which separates the class labels

In case of nonlinear, SVM kernel functions are used to transform the data into different linear separable space. The optimization problem for calculating the hyperplane

$$\min \varphi(w) = \frac{1}{2}(w \cdot w) + C \sum_i \xi_i \tag{43}$$

$$y_i \left(\left(w \cdot \emptyset(x_i) \right) + b \right) \geq 1, \quad \forall i$$

$$C(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j K\left(x_i, x_j\right) \tag{44}$$

Subject to $$\sum_i \alpha_i y_i = 0 \; 0 \leq \alpha_i \leq C \; \forall i.$$

An SVM classifier gives more accurate results and is robust to noise. It also avoids the over fitting problem in training sample space.

Chapter 8

Experimental Results

We conduct our experiments on the Weizmann dataset which is given by Blank et al [16]. This dataset contains the 90 videos with frame rate 15fps and each frame has size 144× 180. In video sequence 9 people performing 10 different actions which are 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.

Figure 8.1: Example of Sample frames from Weizmann Human Action dataset.

In order to find out the accuracy we used the SVM as a multi-class with leave-one-out scheme. In this approach first, we represent the N-Classes dataset into N- two class cases. SVM models are constructed by training all the classes representing positive and negative labels. When test sample is applied classifier compares it with test models and gives voting for winning a class. We use 8 out of 9 people for training purpose and 9^{th} person is utilized for testing. This process continues for all 9 persons in the class and the resulting recognition rates are then averaged.

TABLE -1Confusion Matrix Of MHI and Radon Transform with SVM

	BEND	JACK	JUMP	PJUMP	RUN	SIDE	SKIP	WALK	WAVE 1	WAVE 2
BEND	**9**	0	0	0	0	0	0	0	0	0
JACK	0	**9**	0	0	0	0	0	0	0	0
JUMP	0	1	**8**	0	0	0	0	0	0	0
PJUMP	0	0	1	**8**	0	0	0	0	0	0
RUN	0	0	0	0	**6**	0	1	2	0	0
SIDE	0	0	0	0	0	**8**	0	0	0	1
SKIP	0	0	0	0	2	0	**7**	0	0	0
WALK	0	0	0	0	1	1	0	**7**	0	0
WAVE 1	0	0	0	0	0	0	0	0	**9**	0
WAVE 2	0	1	0	0	0	0	0	0	0	**8**

RECOGNITION RATE – 87.7%

TABLE-2 Confusion Matrix Of MEI and Radon Transform with SVM

	BEND	JACK	JUMP	PJUMP	RUN	SIDE	SKIP	WALK	WAVE 1	WAVE 2
BEND	**9**	0	0	0	0	0	0	0	0	0
JACK	0	**9**	0	0	0	0	0	0	0	0
JUMP	0	1	**8**	0	0	0	0	0	0	0
PJUMP	0	0	1	**8**	0	0	0	0	0	0
RUN	0	0	0	0	**6**	0	1	2	0	0
SIDE	0	0	0	0	0	**8**	0	0	0	1
SKIP	0	0	0	0	2	0	**7**	0	0	0
WALK	0	0	0	0	1	1	0	**7**	0	0
WAVE 1	0	0	0	0	0	0	0	0	**9**	0
WAVE 2	0	0	0	0	0	0	0	0	0	**9**

RECOGNITION RATE -90%

TABLE-3 Confusion Matrix Of AEI and Radon Transform with SVM

	BEND	JACK	JUMP	PJUMP	RUN	SIDE	SKIP	WALK	WAVE 1	WAVE 2
BEND	**9**	0	0	0	0	0	0	0	0	0
JACK	0	**9**	0	0	0	0	0	0	0	0
JUMP	0	1	**8**	0	0	0	0	0	0	0
PJUMP	0	0	0	**8**	0	0	1	0	0	0
RUN	0	0	0	0	**8**	0	0	1	0	0
SIDE	0	0	0	0	0	**9**	0	0	0	0
SKIP	0	0	0	0	2	0	**7**	0	0	0
WALK	0	0	0	0	1	0	0	**8**	0	0
WAVE 1	0	0	0	0	0	0	0	0	**9**	0
WAVE 2	0	1	0	0	0	0	0	0	0	**8**

RECOGNITION RATE -92.22%

TABLE-4 Confusion Matrix Of Still Image and Radon Transform with SVM

	BEND	JACK	JUMP	PJUMP	RUN	SIDE	SKIP	WALK	WAVE 1	WAVE 2
BEND	**9**	0	0	0	0	0	0	0	0	0
JACK	0	**9**	0	0	0	0	0	0	0	0
JUMP	0	0	**9**	0	0	0	0	0	0	0
PJUMP	0	0	0	**9**	0	0	0	0	0	0
RUN	0	0	0	0	**8**	0	0	1	0	0
SIDE	0	0	0	0	0	**8**	0	0	0	1
SKIP	0	0	0	0	2	0	**7**	0	0	0
WALK	0	0	0	0	1	0	0	**8**	0	0
WAVE 1	0	0	0	0	0	0	0	0	**9**	0
WAV E 2	0	1	0	0	0	0	0	0	0	**8**

RECOGNITION RATE -93.35%

TABLE-5 Confusion Matrix Of Still Image and Radon Transform with KNN										
	BEND	JACK	JUMP	PJUMP	RUN	SIDE	SKIP	WALK	WAVE 1	WAVE 2
BEND	**9**	0	0	0	0	0	0	0	0	0
JACK	0	**9**	0	0	0	0	0	0	0	0
JUMP	0	0	**9**	0	0	0	0	0	0	0
PJUMP	0	0	1	**8**	0	0	0	0	0	0
RUN	0	0	0	0	**6**	0	1	2	0	0
SIDE	0	0	0	0	0	**8**	0	0	0	1
SKIP	0	0	0	0	2	0	**7**	0	0	0
WALK	0	0	0	0	1	1	0	**7**	0	0
WAVE 1	0	0	0	0	0	0	0	0	**9**	0
WAVE 2	0	1	0	0	0	0	0	0	0	**8**
RECOGNITION RATE -88.88%										

From the above results of confusion matrix we find that our model best fit with the still images. Recognition rate of AEI further improves if we get good representation of action energy image. The low accuracy in case of MHI and MEI is because of the static posture representation limitation of similar activities. Radon features gives angular variation and helps in improving the results. We also find the spatial distribution plays an important in recognition rate. If we have better distribution around ROI then our results further improves. The effect of pixels mean value is not so high compared to PHOG and Radon but the directional variations give the idea about the representation of posture. In the end we can say that both Radon and PHOG vector give us satisfactory discriminating features to recognize the human action accurately.

Chapter 9

Conclusion and Future work

In this master's project, two approaches, appearance and motion, were incorporated to form Human Action Recognition system. MHI/MEI, AEI and Still images were employed for appearance model, whereas Radon transform was used as feature vector for motion representation.

The study depicted that MHI/MEI images do not improve the recognition as compared to still images. This is because of the reason that it does not give good posture representation for similar type of the activities (running or walking). AEI images, which are dependent on the pixel intensity values, provided better resolution and improved results when compared to MHI/MEI images. However, it is dependent on the alignment of the silhouette and background subtraction model.

For the posture representation, spatial distributions of edge gradients, i.e., PHOG vector, were used. Different spatial distributions are described by PHOG for different activities. As the amount of levels increase in the PHOG vector, we get better-quality spatial distribution and accurate results, but the complexity of the system increases cause of the increase in number of vectors. The characteristics of spatial distribution are dependent on the Region of Interest. The enhanced ROI, i.e. the region with minimal noise, when estimated in the image can bestow improve spatial distribution.

The properties of the Radom transform were embedded with PHOG vector. This method gave improved accuracy in action recognition systems. Radon transform performs on the change in angular variations. For instance, in activities like running, walking and bending, due to variant joint movements, distinctive angular variations are acquired.

This integrated technique supply numerous distinctive feature vectors, which escort us to a robust and noise free action recognition model. In addition, the proposed method is invariant to illumination variations. Nonetheless, this integrated technique can manage to work on a single human activity.

In the future, we would like to work on complex and varied data sets to further check the robustness of the system and to acquire better and more accurate results.

Bibliography

[1] W. Li, Z Zhang, Z. Liu, Action Recognition based on a Bag of 3d points. CVPR workshop for Human communicative Behavior Analysis. 2010.

[2] M. Raptis, D. Kirovski, H. Hoppe, Real-time classification of dance gestures from skeleton animation, Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation – SCA, p. 147, ACM Press, 2011

[3] X. Yang, C. Zhang, Y. Tian, Recognizing actions using Depth Motion Maps based Histograms of Oriented Gradients, Proceedings of the 20th ACM International Conference on Multimedia, MM '12, ACM, New York, NY, USA, pp. 1057–1060, 2012.

[4] J. Wang, Z. Liu, Y. Wu, J. Yuan, Mining actionlet ensemble for action recognition with depth cameras, IEEE Conference on Computer Vision and Pattern Recognition, 2012.

[5] K.N. Tran, I.A. Kakadiaris, S.K. Shah, Part-based motion descriptor image for Human action recognition, Pattern Recognition, 2012.

[6] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, M. F.M. Campos, On the improvement of human action recognition from depth map sequences using Space–Time Occupancy Patterns, Pattern Recognition Letters 36, 221–227, 2014.

[7] J. Luo, W. Wang, H. Qi , Spatio-temporal feature extraction and representation for RGB-D human action recognition, Pattern Recognition Letters ,2014.

[8] J. Wang, Z. Liu,Y. Wu, J. Yuan, Learning Actionlet Ensemble for 3D Human Action Recognition, IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 36, no. 5, 2014.

[9] D.G. Lowe, Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60, 91–110, 2004.

[10] Dalal, N.;Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–26 June 2005; Volume 1, pp. 886–893.

[11] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Transaction on pattern analysis and machine intelligence, vol. 32, no. 9, pp. 1627–1645, 2010.

[12] Mathieu ,Saïda , Boubakeur , Erwan . Ongoing human action recognition with motion capture. Pattern Recognition, 2014, 47, 238–247.

[13] L. Onofri, P. Soda, G. Iannello, Multiple subsequence combination in human action recognition, IET Computer vision, Vol. 8, Issue. 1, pp. 26–34, 2014.

[14] I. Laptev, On space-time interest points, International Journal of Computer Vision, vol. 64 (2-3), 2005.

[15] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, Behavior recognition via sparse spatio-temporal features," in Proc. of the International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005.

[16] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, Actions as Space-Time Shapes, IEEE Transaction. Pattern Analysis and Machine Intelligence, vol. 29, no. 12, pp. 2247-2253, Dec. 2007.

[17] I. Laptev, M. Marszałek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Conference on Computer Vision & Pattern Recognition (CVPR), IEEE, 2008.

[18] J. Niebles, H. Wang, and L. Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words, International Journal of Computer Vision, vol. 79, issue 3, pp. 299-318, 2008.

[19] A. Kovashka, K. Grauman, Learning a hierarchy of discriminative space-time neighborhood features for human action recognition Computer Vision and Pattern Recognition (CVPR), IEEE, 2010.

[20] D. Zhao, L. Shao, X. Zhen, Y. Liu, Combining appearance and structural features for human action recognition, Neurocomputing 113, 88–96, 2013.

[21] Y. Yi, Y. Lin, Human action recognition with salient trajectories, Signal Processing 93 (2013) 2932–2941

[22] X. Zhen, L. Shao, X. Li, Action recognition by Spatio-temporal oriented energies, Information Sciences 281, 295–309, 2014.

[23] G. Somasundaram, A. Cherian, V. Morellas, N. Papanikolopoulos, Action recognition using global spatio-temporal features derived from sparse representations, Computer Vision and Image Understanding 123, 1–13, 2014.

[24] L. Shao, X. Zhen, Spatio-Temporal Laplacian Pyramid Coding for Action Recognition, IEEE Transaction on cybernetics, vol. 44, no 6, 2014.

[25] K.N. Tran, I.A. Kakadiaris, S.K. Shah, Part-based motion descriptor image for human action recognition, Pattern Recognition 45, 2562–2572, 2012.

[26] M. Bregonzio, T. Xiang, S. Gong,Fusing appearance and distribution information of interest points for action recognition, Pattern Recognition 45 , 1220–1234, 2012.

[27] A. Bobick and J. Davis,The Recognition of Human Movement Using Temporal Templates," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 23, no. 3, pp. 257-267, Mar. 2001.

[28] A. Efros, A. Berg, G. Mori, and J. Malik, Recognizing Action at a Distance, Proceeding Ninth IEEE International Conference Computer Vision (ICCV), pp. 726-733, 2003.

[29] M. Roh, H. Shin, S. Lee, View-independent human action recognition with Volume Motion Template on single stereo camera, Pattern Recognition Letters 31, 639–647 ,2010.

[30] L. Shao, L. Ji, Y. Liu, J. Zhang, Human action segmentation and recognition via motion and shape analysis, Pattern Recognition Letters 33 ,438–445, 2012.

[31] G. Goudelis, K. Karpouzis, S. Kollias, Exploring trace transform for robust human action recognition, Pattern Recognition 46 3238–3248, 2013.

[32] C. P. Lee, A. W.C. Tan S. Tan, Time-sliced averaged motion history image for gait recognition, J. Vis. Commun. Image R. 25, 822–826, 2014.

[33] W. Yang, Y. Wang, and G. Mori, Recognizing Human Actions from Still Images with Latent Poses, Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 2030-2037, 2010.

[34] Y. Yang and D. Ramanan, Articulated Pose Estimation with Flexible Mixtures-of-Parts, Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2011.

[35] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, Pose Search: Retrieving People Using their Pose, Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2009.

[36] S. Johnson and M. Everingham, Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation, Proc. British Machine Vision Conf. (BMVC), 2010.

[37] E. Shechtman and M. Irani, Space-time behavioral correlation, in Proc. of the conference on computer vision and Pattern Recognition, 2005.

[38] S. Kumari, S.K. Mitra, Human Action Recognition Using DFT. In Proceedings of the third IEEE National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), Hubli, India, 15–17 December 2011; pp. 239–242.

[39] L. Liu, L. Shao, P. Rockett, Boosted key-frame selection and correlated pyramidal motion-feature representation for human action recognition, Pattern Recognition 46 1810–1818, 2013.

[40] H. Qian, Y. Mao, W. Xiang, Z. Wang, Recognition of Human Activities Using SVM Multi-Class Classifier, Pattern Recognition Letters 31, 100–111, 2010.

[41] A. Sharma , D.K. Kumar, S. Kumar, N. McLachlan, Recognition of Human Actions Using Moment Based Features and Artificial Neural Networks, International Conference on Multimedia Modelling, IEEE, Jan, 2004.

[42] D. Wu, L. Shao, Silhouette Analysis-Based Action Recognition Via Exploiting Human Poses, Transactions on circuits and system for video technology, vol. 23, no. 2, Feb, 2013.

[43] V. H. Chandrashekhar, K. S. Venkatesh, Action Energy Images for Reliable Human Action Recognition, Proc. of ASID , 8-12 Oct, 2006.

[44] J. Han and B. Bhanu, Individual Recognition Using Gait Energy Image, Transactions on Pattern Analysis and Machine Intelligence, IEEE, 316– 322, Feb, 2006.

[45] A. Bosch, A. Zisserman, X. Munoz, Representing Shape with a Spatial Pyramid Kernel, in Proceedings of the ACM International Conference on Image and Video Retrieval, 2007.

[46] A. Bosch, A. Zisserman, X. Munoz, Image Classification using Random Forests and Ferns, ICCV, IEEE, 2007.

[47] W. Kim, Human Action Recognition Using Ordinal Measure of Accumulated Motion, Eurasip Journal on Advances in Signal Processing, 2010.

[48] V. Thanikachalam, K. K. Thyagharatan, Human Action Recognition using Accumulated Motion and Gradient of Motion from Video, ICCCNT, IEEE-20180, July 2012.

[49] M. Singh, M. Mandal, A. Basu, Pose Recognition using the Radon Transform, Circuits and Systems, Vol. 2 , 1091 - 1094 ,2005.

[50] Y. Wang, K. Huang, T. Tan, Human Activity Recognition based on $\mathcal{R}$ Transform Computer Vision and Pattern Recognition, pp. 3722-3729, 2007.

[51] N. V. Boulgouris, Z. X. Chi, Gait Recognition using Radon Transform and Linear Discriminant Analysis, IEEE Transaction on Image Processing, vol. 16, no. 3, March 2007.

[52] J. Li, S. K. Zhou, R. Chellappa, Appearance Modeling using a Geometric Transform, IEEE Transaction on image processing, vol. 18, no. 4, April 2009.

[53] H. Zhang, Z. Liu, H. Zhao, Recognizing Human Activities by Key Frame in Video Sequence, Journal of Software, Vol.5, NO. 8, Aug 2010.

[54] K. Moustakas, D. Tzovaras, G. Stavropoulos, Gait Recognition using Geometric Features and Soft Biometrics, IEEE Signal processing letters vol. 17 no. 4, April 2010.

[55] Z. A. Khan, W. Sohn, Abnormal Human Activity Recognition System based on R-Transform and Kernel Discriminant Technique for Elderly Home Care, IEEE Transactions on Consumer Electronics, Vol. 57, No. 4, November 2011.

[56] A. Jalal, Md. Zia Uddin, T. S. 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, August 2012.

[57] http://www.mathworks.com/help/images/ref/radon.html.

[58] S. Tabbone, L. Wendling, J.P. Salmon, A new shape descriptor defined on the Radon transform, Computer Vision and Image Understanding 102, 42–51, 2006.

[59] Z. A. Khan, W. Sohn, Hierarchical human activity recognition system based on R-transform and nonlinear kernel discriminant features, Electronics Letters, vol: 48, issue: 18, Aug 2012.

[60] L.J.P. Maaten, E.O. Postma, H.J Herik, Dimensionality reduction: A Comparative Review, Online Preprint, 2008.

[61] I.T. Jolliffe, Principal Component Analysis, 2nd edition, New York, Springer (2003).

[62] F. Zheng, L. Shao, Z. Song, X. Chen, Action Recognition using Graph Embedding and the Co-occurrence Matrices Descriptor, International Journal of Computer Mathematics Vol. 88, no. 18, pp. 3896-3914, Dec, 2011.

[63] Chao, W. Lun, "Dimensionality Reduction." Disp Lab, Graduate Institute of Communication Engineering, NTU.

[64] H. Yin, W. Huang, Adaptive nonlinear manifolds and their applications to pattern recognition, Information Sciences 180 , 2649–2662, 2010.

[65] S. Roweis, L. Saul, Nonlinear Dimensionality reduction by locally linear embedding, Science 290 (5500), 2323–2326, 2000.

[66] J. Chen, Y. Liu, Locally linear embedding: A survey, Artificial intelligence Rev, 36:29–48, 2011.

[67] Y. Huang, Discriminant diffusion maps analysis: A robust manifold learner for dimensionality reduction and its applications in machine condition monitoring and fault diagnosis, Mechanical Systems and Signal Processing 34 (2013) 277–297.

[68] Y. Wang, H. Jiang, M. Drew, Z. N. Li, G. Mori, Unsupervised discovery of action classes, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2006, pp. 1654–1661.

[69] P. Li, J. Ma, What is happening in a still picture?, in: IEEE Asian Conference on Pattern Recognition, IEEE, 2011, pp. 32–36.

[70] B. Yao, A. Khosla, L. Fei-Fei, Combining randomization and discrimination for fine-grained image categorization, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2011, pp. 1577–1584.

[71] B. Yao, X. Jiang, A. Khosla, A. Lin, L. Guibas, L. Fei-Fei, Human action recognition by learning bases of action attributes and parts, in: IEEE International Conference on Computer Vision, IEEE, 2011, pp. 1331–1338.

[72] L.-J. Li, L. Fei-Fei, What, where and who? classifying events by scene and object recognition, in: IEEE Conference on Computer Vision, IEEE, 2007, pp. 1–8.

[73] C. Thurau, V. Hlavac, Pose primitive based human action recognition in videos or still images, in: IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–8.

[74] A. Lopes, E. Santos, E. Valle, J. Almeida, A. Araujo, Transfer Learning for Human Action Recognition, SIBGRAPI Conference on Graphics, Patterns and Images, 2011.

[75] Y. Zheng, Y. Zhang, X. Li, B. Liu, Action Recognition in still images using a combination of human pose and context information, International Conference on Image Processing (ICIP), IEEE, 2012.

[76] J. Hu, W. Zheng, J. Li, S. Gong, T. Xiang, Recognizing Human-Object Interaction via Exemplar based Modelling, International Conference on Computer Vision, IEEE, 2013.

[77] G. Sharma, F. Jurie, C. Schmid, Expanded Parts Model for Human Attribute and Action Recognition in Still Images, Conference on Computer Vision and Pattern Recognition, IEEE, 2013.

[78] A. A. Chaaraoui, P. C. Perez, F. F. Revuelta, Silhouette-based human action recognition using sequences of key poses, Pattern Recognition Letters 34,2013, pp. 1799–1807.

[79] H. Zhang, J. Wu, D. Zhong, S, W. Smoliar, An integrated system for content-based video retrieval and browsing, Pattern Recognition, vol. 30, no. 4, 1997, pp. 643-658.

[80] S. Mehrotra, Adaptive key frame extraction using unsupervised clustering, Proc. International Conference on Image Processing, 1998.

[81] P. Zeng, Z. Chen, Perceptual quality measure using JND model of the human visual system, 2011 International Conference on Electric Information and Control Engineering, 2011.

[82] J. C. Burges, A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discover 2 (2): 121–167, 1998.