USE OF PROBABILITY HYPOTHESIS DENSITY FILTER FOR HUMAN ACTIVITY RECOGNITION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ELİF ERDEM GÜNAY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

MARCH 2016

Approval of the thesis:

USE OF PROBABILITY HYPOTHESIS DENSITY FILTER FOR HUMAN ACTIVITY RECOGNITION

submitted by **ELİF ERDEM GÜNAY** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Electrical and Electronics Engineering Department, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Gönül Turhan Sayan
Head of Department, **Electrical and Electronics Eng.** _____

Prof. Dr. Gözde Bozdağı Akar
Supervisor, **Electrical and Electronics Eng. Dept., METU** _____

Prof. Dr. Mübeccel Demirekler
Co-supervisor, **Electrical and Electronics Eng. Dept., METU** _____

Examining Committee Members:

Prof. Dr. Orhan Arıkan
Electrical and Electronics Eng. Dept., Bilkent University _____

Prof. Dr. Gözde Bozdağı Akar
Electrical and Electronics Eng. Dept., METU _____

Prof. Dr. Aydın Alatan
Electrical and Electronics Eng. Dept., METU _____

Assoc. Prof. Dr. Umut Orguner
Electrical and Electronics Eng. Dept., METU _____

Assist. Prof. Dr. Sevinç Figen Öktem
Electrical and Electronics Eng. Dept., METU _____

Date: _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: ELİF ERDEM GÜNAY

Signature :

ABSTRACT

USE OF PROBABILITY HYPOTHESIS DENSITY FILTER FOR HUMAN ACTIVITY RECOGNITION

GÜNAY, ELİF ERDEM
Ph.D., Department of Electrical and Electronics Engineering
Supervisor : Prof. Dr. Gözde Bozdağı Akar
Co-Supervisor : Prof. Dr. Mübeccel Demirekler

March 2016, 140 pages

This thesis addresses a Gaussian Mixture Probability Hypothesis Density (GM-PHD) based probabilistic group tracking approach to human action recognition problem. First of all, feature set of the video images denoted as observations are obtained by applying Harris Corner Detector(HCD) technique following a GM-PHD filter, which is a state-of-the-art target tracking method. Discriminative information is extracted from the output of the GM-PHD filter and using these, recognition features are constructed related to different body segments and the whole body. An unique Hidden Markov Model(HMM) belonging to each feature is fed by these information and recognition is performed by selecting optimal HMM's. The performance of the proposed approach is shown on the videos in KTH Research Project Database and custom videos including occlusion scenarios.The results are presented as the percentage of the correctly recognized videos. Same experiments on KTH database are performed for KLT tracker instead of GMPHD in the proposed approach. In addition, a comparison is made

for an algorithm in the literature for the custom videos. The results shown that proposed approach has comparable performance on KTH database and is better in handling occlusion scenarios.

Keywords: Human action recognition, Gaussian Mixture PHD, Hidden Markov Model, Harris Corner Detector

ÖZ

İNSAN HAREKET ALGILAMA İÇİN OLASILIKSAL HİPOTEZ YOĞUNLUK FİLTRESİ KULLANIMI

GÜNAY, ELİF ERDEM
Doktora, Elektrik ve Elektronik Mühendisliği Bölümü
Tez Yöneticisi : Prof. Dr. Gözde Bozdağı Akar
Ortak Tez Yöneticisi : Prof. Dr. Mübeccel Demirekler

Mart 2016 , 140 sayfa

Bu tez, GMPHD tabanlı olasılıksal grup takibi ile insan hareketi algılama konusunu adreslemektedir. Önerilen çözüm, probleme sıralı ve durumsal bir şekilde yaklaşmaya dayanmaktadır. Öncelikle, Harris Corner Detector(HCD) tekniği kullanılarak video görüntülerindeki özellik setleri elde edilmektedir. Bunu takiben elde edilen özellik kümeleri üzerinde son yıllarda popüler olan Gaussian Mixture Probability Hypothesis Density (GMPHD) filtresi uygulanmaktadır. GMPHD filtresi çıktısından farklılık yaratan bilgiler çıkarılmakta ve bu bilgiler ile hem insan vücudunun farklı bölümleri ile ilişkili hem de tüm vücudu temsil eden algılama özellikleri çıkarılmaktadır. Bu özellikler her bir özellik için oluşturulan Hidden Markov Modellerini(HMM) beslemekte ve optimal HMM'lerin seçimi ile algılama sonucu oluşturulmaktadır. Önerilen çözümün başarımı KTH Araştırma Projeleri Veri Tabanı ve tez kapsamında çekilmiş insanın sabit bir objenin arkasından geçtiği durumlardaki videolarda gösterilmiş ve sonuçlar doğru

bir şekilde tanımlanmış videoların yüzdesi olarak sunulmaktadır. Aynı deneyler KTH veritabanında, önerilen algoritmanın takip adımında KLT kullanılarak da gerçekleştirilmiş ve tanıma yüzdeleri çıkarılmıştır. Ayrıca literatürde bulunan yüksek performanslı bir algoritma ile bu videolar üzerinden karşılaştırma yapılmıştır.

Anahtar Kelimeler: insan hareketi algılama, Gaussian Mixture PHD, Hidden Markov Model, Harris Corner Detector

To my beloved family...

ACKNOWLEDGMENTS

I would like to thank my supervisor Professor Akar for his constant support, guidance and friendship. It was a great honor to work with her for the last ten years and our cooperation influenced my academical and world view highly. I also would like to thank Prof. Demirekler for her strong support, guidance and patience during all phases of this thesis study.

This thesis is also supported by my company ASELSAN by allowing me to focus on my Ph.D. study and realizing the importance of combining the real world engineering problems with the academic approaches, in an appropriate way.

My parents also has provided invaluable support for this work. I would like to thank specially to my mother Gülbeyaz for her life-energy and my Father Mümin for his humility and simplicity who always make me feel loved and cared. I also want to thank to my mother-in law Vildan and my father-in law Mehmet for their generosity, love and equanimity.

Last words goes to my priceless nuclear family. My first thanks go to my husband Melih. You always make me feel special and loved, you are my life partner. Secondly, I would like to thank to my beauty Ada. She is my flower of love and life energy. I also would like to thank to my little baby. You and Ada are the source of my happiness and meaning of life. Me and your father do our best in order to raise you as your own. I wish you a happy life my dears. This thesis would not appear in this way without their endless love...

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

AHMM	Abstract Hidden Markov Model
AMI	Accumulated Motion Image
CCA	Canonical Correlation Analysis
CHSMM	Coupled Hidden Semi-Markov Model
DBN	Dynamic Bayesian Networks
FISST	Finite Set Statistics
GEI	Gait Energy Image
GMPHD	Gaussian Mixture Probability Hypothesis Density
GR-1	Group-1
GR-2	Group-2
HAR	Human Action Recognition
HCD	Harris Corner Detection
HHMM	Hierarchical Hidden Markov Model
HMM	Hidden Markov Model
HoF	Histogram of Optical Flow
HoG	Histogram of Gradient
HLR	High Level Recognition
HSMM	Hidden Semi-Markov Model
JPDA	Joint Probabilistic Data Association
KF	Kalman Filter
KLT	Kanade-Lucas-Tomasi
LP	Lower Part
MBH	Motion Boundary Histogram
MHT	Multiple Hypothesis Tracking
MIL	Multiple Instance Learning
NNC	Nearest Neighborhood Classification
PF	Particle Filter
PCA	Principle Component Analysis

PDA	Probabilistic Data Association
PHD	Probability Hypothesis Density
PMHT	Probabilistic Multiple Hypothesis Tracking
RFS	Random Finite Sets
SVM	Support Vector Machine
UL	Upper Left
UP	Upper Part
UR	Upper Right

CHAPTER 1

INTRODUCTION

1.1 Introduction

Human Action Recognition (HAR) is a complex problem which requires individual solutions to detection, tracking and recognition sub-problems. This thesis is an initial work showing the usability of GMPHD filter for tracking phase in HAR problems. This filter is a promising tool for multi-target tracking problems and is capable of group tracking when the number of targets in the scene is changing as in the occlusion case. In the detection step, features are extracted by Harris Corner Detector and then tracked by GMPHD filtering technique as a group, which is a state-of-the-art multi-target tracker. Afterwards, we utilize patterns extracted from GMPHD filter intensity and use these patterns to recognize the actions by utilizing Hidden Markov Models (HMM).

In the literature, there is no existing solution which utilizes GMPHD filter for human motion analysis.The underlying idea is that if the body in the scene is represented by the composition of multiple identical type features, the filter can handle the varying number of features and can group track all the features instantaneously.

In video application problems, number of features is very likely to change for each frame because of the following parameters:

- articulated parts of the body such as arms and legs resulting in occlusion of the features,

- change in the background scene and
- dynamic noise in the scene

GM-PHD filter is able to continuously track the occluded features for a certain period of time by its inner mechanism which make it faster to include occlusion to group tracking. Birth mechanism also helps capture new measurements to group tracking in a short period of time.

Apart from the change of the number of features in the video sequence, GM-PHD filter performs the association and tracking at the same level which severely reduces the computation and the complexity of the solution.

By applying GM-PHD filter to a video sequence, 2D image information is converted to a 3D intensity function which is a Gaussian Mixture. In this domain, first two dimensions correspond to the positional information while 3rd dimension can be taken as the weights of the Gaussians in the mixture. Note that the sum of weights corresponds to expected number of features in the scene. States of GM-PHD includes information about position and velocity and represents the dynamical behavior of the body parts like torso, arms, legs and head. To be able to track the human body, all these articulated and non-articulated parts should be tracked individually. In this thesis, GM-PHD constant velocity motion model is considered. Since there are different types of action with different motion characteristics like sinusoidal as in walking or linear as in handclapping, linear constant velocity is assumed to be the common subset of these motions.

As stated in chapter 3, proposed solutions for human activity recognition covers several different image processing techniques and evaluated on several different image databases. These studies generally focuses on extracting the information from the image intensity and individual tracks of these features in time. These solutions declares quite well performances up to the correct recognition of 95% of the videos in the related databases as given in [1] and in section 3.

In this thesis, our aim is to model the human body parts as multiple-targets and apply GM-PHD filter, a multi-target tracker, to this problem. As another novelty, we utilize image intensity information to guide GM-PHD filter in which

difference between the image patches around the measurement is utilized in weight determination of GM-PHD. The features are extracted from the image intensities as the measurements to the tracker, yet the recognition step is based only on the output of the multi-target GM-PHD tracker. The recognition parameters are extracted from the Gaussian Mixture intensity function belonging to the whole body and different parts of the body, as well. The recognition phase, including both learning and testing, is based on the HMM technique and the recognition parameters are certainly translated to the language of an HMM structure. Explanations of these parameters and required pre-processing steps are provided chapter 6.

1.2 Problem Definition-Overview

The aim of the thesis is to identify the action taken by the human in a given video sequence. The video set is selected as the KTH database described in [60], which is a controlled database composed of videos of 25 individuals performing 6 different action (walking, running, jogging, boxing, hand-clapping and hand-waving) in 4 different environments. Detailed information about the data base is given in Chapter 6.

KTH database is chosen since it includes diverse environments when compared to others. When the actions in the videos of KTH are analyzed, they seem to have different characteristics which bring difficulties to the recognition problem.

1. Different aspects of the body (i.e., front, back and side pose)
2. Body with different clothes and properties (i.e., gloves, topcoat, long haired)
3. Zooming-in and out operations
4. Existence of Shadows in different angles
5. Actions towards different directions.

In order to eliminate these undesired effects on the recognition process, all the information should be brought to a common reference. So the elimination of the information belonging to the background scene but not the body itself should be performed and mapping of the feature parameters to the reference point should be done.

To speak specifically for each problem, if the aspect of the body is different, then the feature distribution of the body will be different which affects recognition problem significantly. For this problem, we extract the height of the body in action and normalize velocity related features using height information. But considering side, front and back pose of the body, the difference in feature distribution is still a challenge for our problem. Besides, the individuals performing action wear different clothes. As examples, the person wearing topcoat and has long hair affect the motion characteristics of the body significantly, gloves cause poor feature extraction at hand in darker background etc.

There is also zooming operation in the videos which causes;

1. Change in velocity information
2. Change in height

In order to decrease the negative effects of zooming, we normalize velocity vectors using the height information of the body. But the virtual velocity in X and Y direction still bring difficulty in recognition. In order to decrease the influence of shadows, we decrease the number of features around shadows, but this operation decreases the number of features on the body. Considering the actions in difference direction, we bring all motions to the same direction as given in 5.3. The operations on the features because of not only video adversity but also for recognition purposes are described in the following sections in detail.

1.3 Problem Solution-Overview

In this thesis, we intend to make recognition of human actions in a controlled database using GM-PHD filter output properties and propose a different solution

to this problem. The flow chart of the training and testing phases of the proposal is given in Figure 1.1 and Figure 1.2, respectively.

2D image intensity contains information which does not only belong to body in action but also to the background. So, as the first step, discriminative information in the image has to be extracted. We used Harris Corner Detector technique to extract features belonging to the body. The reason of selection of the Harris Corner Detector as the feature extractor is its robustness and consistency and suitability for multi-target tracking by extracting multiple corners in an image.

Group tracking is performed on the Harris corners by the GM-PHD filter which yields intensity function including the state information of each feature. After elimination of outlier features using estimated set of states, they are re-calculated with respect to the common reference scaling. Then parameters, that will be used in the training phase, are calculated from GM-PHD intensity function. As a final stage, an unique HMM structure for each action is built which are the combination of HMMs belonging to each feature given in tables 6.5 and 6.6. This implies each feature parameter has independent HMMs. After training phase, the testing videos in the database are recognized by the trained HMM structures and the recognition results are obtained by a voting algorithm in an optimal way.

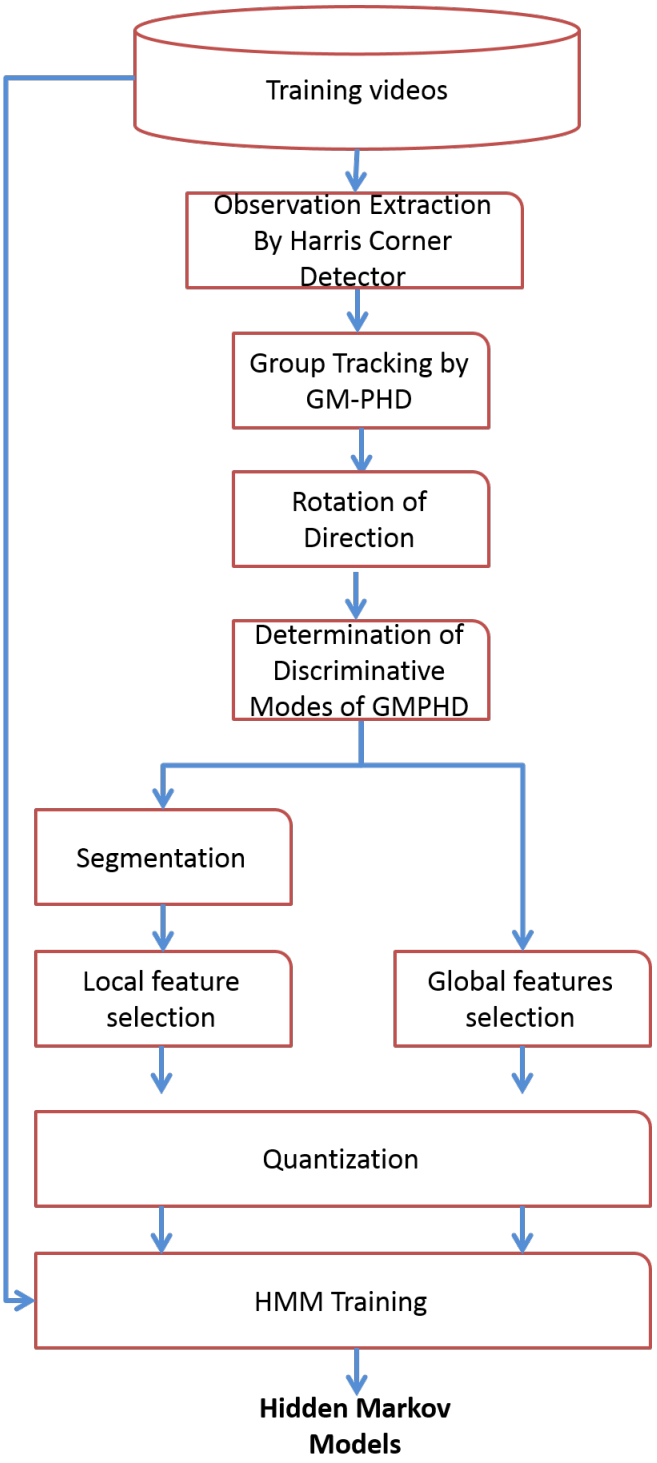


Figure 1.1: Training Flow of the Proposed Algorithm

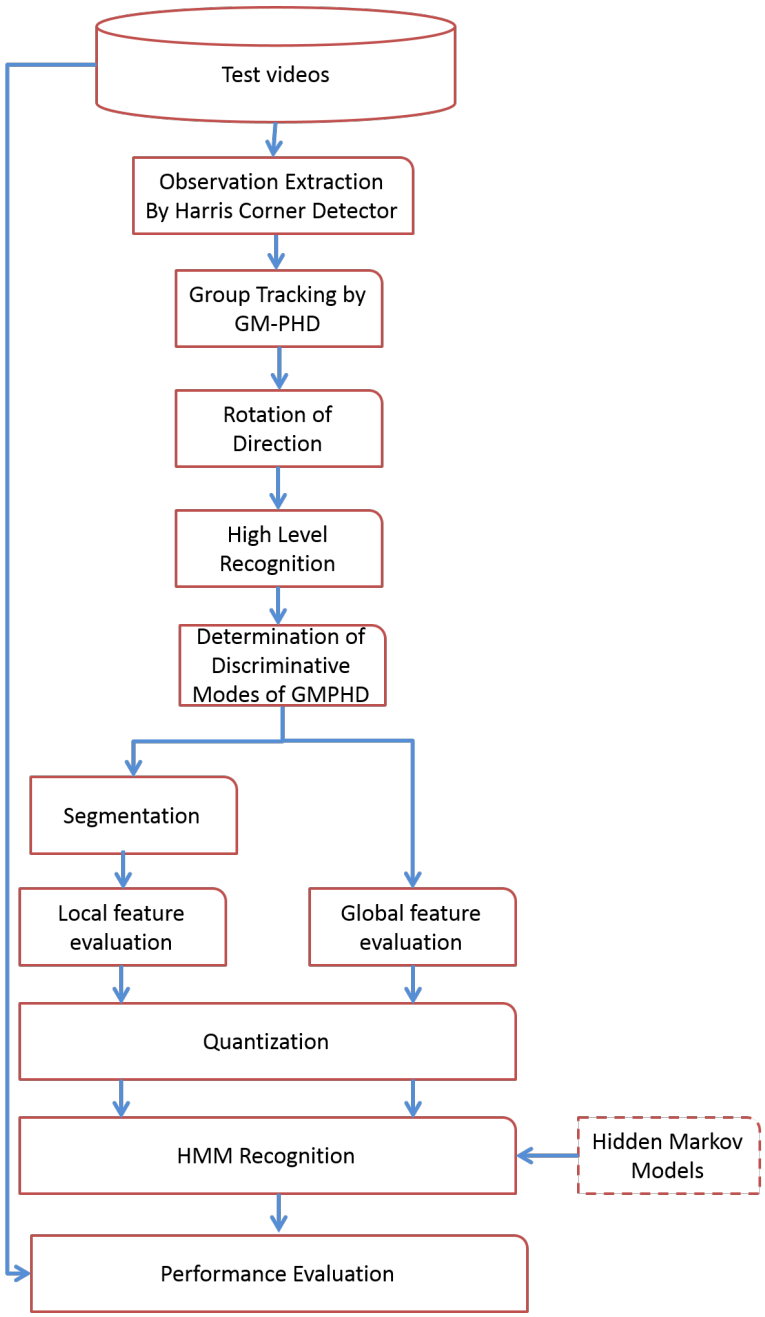


Figure 1.2: Testing Flow of the Proposed Algorithm

1.4 Performance Comparison

The performance of the proposed algorithm is obtained by the videos in a controlled database which gives a chance to compare with the existing algorithms. The performance of the algorithm is about 89% and there are approaches with higher performance results in the literature. In order to reveal the advantage of GMPHD, we took custom videos including occlusion and compared the recognition performances of one of the most successful algorithm [54] and our approach. We realize that our approach is more noise insensitive compared to the [54].

In addition we compared the group tracking performance of the GMPHD Filter with tracking performance of KLT. This is performed by replacing GMPHD with KLT tracker and evaluating the algorithm with the same parameter and video sets. We see that our approach yields 10% more recognition performance than the one with KLT.

CHAPTER 2

BACKGROUND

There are several representations of human body and corresponding tracking methods in the literature. The most important parameter for solving the problem of action recognition is obviously the variety and number of information types extracted from the video sequence. There are both deterministic and probabilistic approaches utilized in HAR problem. Deterministic approaches directly uses the information extracted from the video sequence which does not provide any information related to the statistics of the video. On the other hand statistical approaches extracts and utilize the statistical information in the video sequence. In this thesis both deterministic and statistical information is utilized for action recognition purposes. Feature detection is performed in deterministic ways whereas group tracking utilize the statistical information related to the target dynamics and recognition is performed utilizing the statistical properties.

In HAR occlusion of the features is one of the major problems. It causes big problems in deterministic approaches whereas the used group tracking method, PHD filtering, inherently deals with any type of occlusion within the filter. Apart from information related to the motion model of the features, PHD filtering reveals not only the information related to the motion characteristics ,i.e., position, velocity etc., but also the number of features at each frame.

In this thesis, we propose a complete algorithm which is capable of action recognition in which features are extracted using Harris Corner Detector, group-tracking of the features is performed with GM-PHD. The recognition is mainly based on Hidden Markov Models and after extracting the desired information.

The algorithm is trained and tested by this technique.

Detailed information related to HCD, GM-PHD, and HMM will be provided in this section for the sake of completeness. The novel approach based on these techniques will be explained in Chapter 5.

2.1 Harris Corner Detector

Corners are defined as the intersection point of two edges. So there must be significant change in appearance when we shift the window around the corner in any direction. Harris Corner Detector, [19] uses this idea and gives a mathematical formulation for determining these points.

Define the intensity change when we shift the window by [u,v] as given in Equation (2.1).

$$E(u,v) = \sum_{x,y} w(x,y)[I(x+u, y+v] - I(x,y)]^2 \tag{2.1}$$

where w is windowing function and I is the intensity. Note x,y is the location of the point in the image.

Using the first order approximation of Taylor Series for 2D functions the square term of (2.1) turn into Equation 2.2.

$$[I(x+u, y+v] - I(x,y)]^2 = \begin{bmatrix} u & v \end{bmatrix} \begin{bmatrix} I_x^2 & I_{x,y} \\ I_{x,y} & I_y^2 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \tag{2.2}$$

where I_x and I_y are image derivatives of image I . For small displacements, we apply a bilinear approximation and obtain the Equation (2.3).

$$E(u,v) = \begin{bmatrix} u & v \end{bmatrix} M \begin{bmatrix} u \\ v \end{bmatrix} \tag{2.3}$$

s

$$M = \sum_{x,y} w(x,y) \begin{bmatrix} I_x^2 & I_{x,y} \\ I_{x,y} & I_y^2 \end{bmatrix} \tag{2.4}$$

Analyzing of M matrix gives us the corner points. First define the measure of corner response, R, as given in (2.5).

$$R = detM - k(trace(M))^2 \tag{2.5}$$

where

$$detM = \lambda_1 \lambda_2 \tag{2.6}$$

$$traceM = \lambda_1 + \lambda_2 \tag{2.7}$$

where λ_1 and λ_2 are eigenvalues of M matrix. Note that k is an empirically determined constant generally chosen between $k = 0.04 - 0.06$

Corners are the points which have large R values which corresponds to the M matrix with large and similar eigenvalues.

2.2 PHD Filter and Its Gaussian Mixture Implementation

In [17] and [44], the random set theory is defined as a theoretical framework for multisensor-multitarget data processing in which a set of multiple targets at time t is represented as a Random Finite Set (RFS). The Finite Set Statistics (FISST) is the first systematic treatment of multisource-multitarget problem which transforms the multisource-multitarget problem into a mathematically equivalent single-sensor, single-target problem. The multitarget states and observations are represented as finite sets instead of the vector notation. The representation of the state of a target (commonly chosen as position and velocity) is represented by a state vector x, and the RFS is $X_t = \{x_1, x_2, .., x_{N(t)}\}$ where $N(t)$ is variable target number at time t. Similarly, the measurement set is $Y_t = \{y_1, y_2, .., y_{M(t)}\}$ where $M(t)$ is variable measurement number at time t.

The problem of FISST is its combinatorial complexity in the multi-target case. Mahler proposed Probability Hypothesis Density (PHD) to approximate optimal Bayes Filter and to reduce the complexity in [45]. The idea of PHD filtering is based on Finite Set Statistics (FISST) whose underlying idea is treating finite sets as random elements from probability theory point of view. In PHD framework, the data obtained from various target/source is unified under a single Bayesian framework and all the detection, tracking and identification problems become a single problem.

In Human Tracking applications, two different implementations of PHD filtering has been used. The particle filter or the sequential Monte Carlo method is a Monte Carlo simulation based recursive Bayes filter and can be applied to solve nonlinear and non-Gaussian problems. Other implementation which is called Gaussian Mixture PHD (GM-PHD) is proposed for the linear, Gaussian target dynamic model and birth process by [67] and it brings a closed form solution to iterative calculation of means, covariance matrices and weights of the filter.

In this section, further information for GM-PHD filter is given for better understanding of the usage of the technique in the thesis solution.

Gaussian Mixture PHD (GMPHD) filter brings forward a closed form solution to the multiple target tracking problem. [67] explains the theory and pseudo-code of this method yet the pseudo code of the algorithm is also given in the appendix of the thesis (Section A).

Basic assumptions for the derivation of the filter are:

- target dynamical model for a single target is linear Gaussian :

$$p_{k|k-1}(z|x) = \mathcal{N}(z; A_{k-1}x, Q_{k-1}) \tag{2.8}$$

- measurement model is also linear Gaussian:

$$p_{\tilde{y}|\tilde{x}}(y|x) = \mathcal{N}(y; C_k x, R_k) \tag{2.9}$$

- Target intensity, $D_{\widetilde{X}_k}(x)$, is in the form of a Gaussian mixture:

$$D_{\widetilde{X}_k}(x) = \sum_{i=1}^{J_{\widetilde{X}_k}} w_{\widetilde{X}_k}^{(i)} \mathcal{N}(x; m_{\widetilde{X}_k}^{(i)}, P_{\widetilde{X}_k}^{(i)}) \tag{2.10}$$

- Birth intensity, $D_{\widetilde{B}_k}(x)$, is also a Gaussian mixture:

$$D_{\widetilde{B}_k}(x) = \sum_{i=1}^{J_{\widetilde{B}_k}} w_{\widetilde{B}_k}^{(i)} \mathcal{N}(x; m_{\widetilde{B}_k}^{(i)}, P_{\widetilde{B}_k}^{(i)}) \tag{2.11}$$

Based on the assumptions above, predicted target intensity and measurement updated target intensity are also found as Gaussian mixtures as in (2.12) and (2.13), respectively.

$$\widehat{D}_{\widetilde{X}_{k+1}}(x) = \sum_{j=1}^{J_{\widetilde{B}_{k+1}}} w_{\widetilde{B}_{k+1}}^{(j)} \mathcal{N}(x; m_{\widetilde{B}_{k+1}}^{(j)}, P_{\widetilde{B}_{k+1}}^{(j)}) + p_S \sum_{i=1}^{J_{\widetilde{X}_k}} w_{\widetilde{X}_k}^{(i)} \mathcal{N}(x; \widetilde{m}_{\widetilde{X}_k}^{(i)}, \widetilde{P}_{\widetilde{X}_k}^{(i)}) \tag{2.12}$$

$$\begin{aligned} D_{\widetilde{X}_{k+1}}(x) = (1 - p_D) \sum_{j=1}^{J_p} w_p^{(j)} \mathcal{N}(x; m_p^{(j)}, P_p^{(j)}) + \\ \sum_{j=1}^{J_p} \sum_{y \in Y} \frac{w_p^{(j)} q^{(j)}(y) p_D}{D_{\widetilde{C}}(y) + p_D \sum_{i=1}^{J_p} w_p^{(i)} q^{(j)}(y)} \mathcal{N}(x; \widetilde{m}^{(j)}, \widetilde{P}^{(j)}) \end{aligned} \tag{2.13}$$

Note that pruning operation is required after obtaining the measurement updated PHD to reduce the number of Gaussians to a certain level. After the pruning phase, estimated positions of the targets are calculated by considering the Gaussian mixture weights. Pseudo-code for pruning and state extraction is given by the tables A.3 and A.4 in the appendix. Flow of the algorithm implemented for the thesis is provided in Figure 2.1.

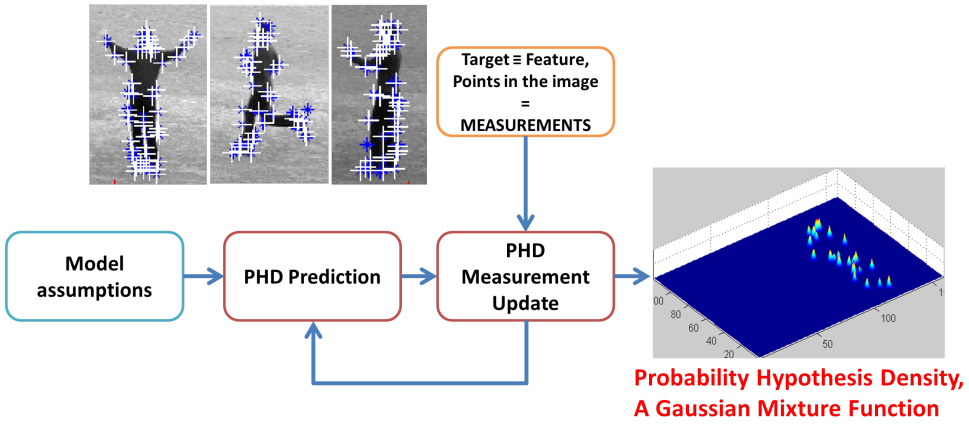


Figure 2.1: High Level Process Flow for the GMPHD filter

2.3 Hidden Markov Models

Hidden Markov Model is a stochastic model which is a widely used probabilistic framework in action recognition. HMMs and their extensions are well suited for action recognition because of its sequential nature. They assume that the underlying signal is of Markov Model which corresponds to that we can estimate the next action from the current one and this is not far beyond our case when we consider human actions visually.

In order to well suit the HMM model to the problem, fundamental model parameters have to be selected properly. The selection of number of hidden states and both the model and number of observation symbols are three of these. In learning phase of HMM, the values of the statistical parameters are determined and in the testing phase the model giving the highest posterior probability is taken as the action. In the next part, a background for HMM is given.

Theory of the Hidden Markov Model

HMM characterizes the statistical properties of the given signal which corresponds to the sequences of feature values belonging to human body/action for HAR case.

An HMM is characterized by the following parameters ([53]):

i N, number of hidden states in the model, S is the set of states and the state variable at time t is denoted by q_t where $q_t = S_j$ means that at time t the state is S_j.

$$S = \{S_1, S_2, ..., S_N\} \tag{2.14}$$

ii M, number of distinct observation symbols per state and V is the symbol set.

$$V = \{v_1, v_2, ..., v_M\} \tag{2.15}$$

If O_t is the observation at time t then $O_t = V_j$ means that the observation at time t is V_j.

iii A is an $N \times N$ probability transition matrix

$$A = [a_{i,j}]_{NxN} \tag{2.16}$$

where $a_{i,j} = P[q_{t+1} = S_j | q_t = S_i]$, $1 \leq i, j \leq N$

iv B denotes the conditional mass function of observations. B is a vector of size M where its j^{th} entry is denoted by b_j is given below

$$B = \{b_j(k)\} \tag{2.17}$$

where $b_j(k) = p[V_k = O_t | q_t = S_j]$, $1 \leq j \leq N, 1 \leq k \leq M$.

v π, the initial state distribution,

$$\pi = [\pi_i]_{1 \times N} \tag{2.18}$$

where $\pi_i = P\{q_1 = S_i\}$, $1 \leq i \leq N$

Then the HMM model is represented as $\lambda = (A, B, \pi)$.

In this thesis we construct a HMM with 4 hidden states and 11 observation symbols. And the transition probability matrix and observation probability distributions are calculated by using the training data set.

Note that the observation symbols correspond to the physical output of the system being modeled which are the quantized values of each feature given in Table 6.5 and 6.6. Then, the observation symbols corresponds to the numbers between 1 and 11.

There are three fundamental problems of HMM.

i Problem 1 : Evaluation Problem

The first problem is the evaluation problem, i.e., calculating the probability that observed sequence is produced by the model. The solution of this problem helps us classify the action in the testing phase.

ii Problem 2 : Decoding Problem

The second is decoding problem which calculated the most likely path of the hidden states given the observation sequence.

iii Problem 3 : Problem of Adjusting Model Parameters (Learning)

The last problem is the calculation of the HMM parameters $\lambda = (A, B, \pi)$ by the training data.

Solution of Evaluation Problem Assume that the model $\lambda = (A, B, \pi)$ and observation sequence $O_1O_2O_3...O_t$ is given. The probability that this sequence is produced by the model i.e., $P(O|\lambda)$ is calculated by forward-backward procedure efficiently. First define the forward variable, $\alpha_t(i)$, that gives the probability of the partial observation sequence until time t and hidden state is $q_t = S_i$.

$$\alpha_t(i) = P(O_1O_2O_3..O_t, q_t = S_i|\lambda) \tag{2.19}$$

$\alpha_t(i)$ can be solved inductively with forward part of forward-backward procedure 2.22.

1. **Initialization:**

$$\alpha_1(i) = \pi b_i(O_1),\ 1 \leq i \leq N \tag{2.20}$$

2. **Induction:**

$$\alpha_{t+1}(j) = \left(\sum_{i=1}^{N} \alpha_t(i) a_{ij} \right) b_j(O_{t+1}),\ 1 \leq t \leq T-1,\ 1 \leq j \leq N \quad (2.21)$$

3. **Termination:**

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i) \quad (2.22)$$

In the initialization step, the forward probability is defined as the joint probability of state S_i and the initial observation of O_1. In induction step, S_j can be reached at time t from N possible states. So summing over all the N possible states S_i and possible transitions a_{ij} at time t results in probability of S_j at time $t+1$. Then $\alpha_{t+1}(j)$ is calculated just by accounting for observation O_{t+1} in state j. In termination, summation over all terminal forward variables $\alpha_T(i)$is performed.

In order give a complete solution backward recursion is presented here which is indeed utilized in the solution of Problem 3.

The induction approach is very similar. First we define a backward variable $\beta_t(i)$ as the probability of the partial observation sequence from $t+1$ to the end.

$$\beta_t(i) = P(O_{t+1}O_{t+2}O_{t+3}..O_T, q_t = S_i|\lambda) \quad (2.23)$$

The way of solving the backward variable, $\beta_t(i)$, is given in 2.26:

1. **Initialization:**

$$\beta_T(i) = 1,\ 1 \leq i \leq N \quad (2.24)$$

2. **Induction:**

$$\beta_t(i) = \left(\sum_{j=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(i) \right), \ t = T-1, T-2, \ldots, 1 \ , \ 1 \leq i \leq N \tag{2.25}$$

3. **Termination:**

$$P(O|\lambda) = \sum_{i=1}^{N} \beta_1(i) b_i(O_1) \pi_i \tag{2.26}$$

Solution of the Decoding Problem

Decoding problem can be stated as an optimization problem which maximizes $P(S|O, \lambda)$ where S is the state sequence. There are different approaches which finds the most likely sequence producing the given observation sequence. The approach we utilize is called Viterbi algorithm which is based on dynamic programming method.

To find the best state sequence, $Q = \{q_1 q_2 .... q_T\}$ for a given observation sequence $O = \{O_1 O_2 .... O_T\}$, we define the quantity

$$\delta_t(i) = max_{q_1 \ldots q_{t-1}} P(q_1 q_2 .... q_t = i, O_1 O_2 .... O_t | \lambda) \tag{2.27}$$

where $\delta_t(i)$ is the highest probability along a single path, at time t where the first t observations ends in state S_i.

The next step at time t+1 is:

$$\delta_{t+1}(j) = [max_i \delta_t(i) a_{ij}] . b_j(O_{t+1}) \tag{2.28}$$

In order to track the max of (2.28), the $\psi_t(j)$ is defined and calculated as:

1. **Initialization:**

$$\delta_1(i) = \pi_i b_i(O_1), \ 1 \leq i \leq N \tag{2.29}$$

$$\psi_1(i) = 0. \tag{2.30}$$

2. **Recursion:**

$$\delta_t(j) = max_{1\leq i\leq N}\ [\delta_{t-1}(i)a_{ij}]b_j(O_t),\ 2 \leq t \leq T,\ 1 \leq j \leq N \tag{2.31}$$

$$\psi_t(j) = argmax_{1\leq i\leq N}\ [\delta_{t-1}(i)a_{ij}],\ 2 \leq t \leq T,\ 1 \leq j \leq N \tag{2.32}$$

$$\tag{2.33}$$

3. **Termination:**

$$P^* = max_{1\leq i\leq N}[\delta_t(i)] \tag{2.34}$$

$$q_T^* = argmax_{1\leq i\leq N}[\delta_t(i)]. \tag{2.35}$$

4. **Backtracking:**

$$q_t^* = \psi_{t+1}(q_{t+1}^*),\ t = T-1, T-2, .., 1. \tag{2.36}$$

Viterbi algorithm is very similar to forward step of evaluation problem but for backtracking step. The maximization in (2.32) over previous steps which is used in (2.21) is the fundamental difference.

Solution of Learning Problem

The most comprehensive and difficult problem of HMM is to determine the model parameters, $\lambda = (A, B, \pi)$, given the data and the model. In fact, there is no analytic and optimal way of estimating the model parameters. On the other hand one can find $\lambda = (A, B, \pi)$ which locally maximizes $P(O|\lambda)$ in an iterative manner. This is the Baum-Welch algorithm which is based on expectation maximization:

First, let's define the probability of being in state S_i at time t and S_j at time t+1 as follows:

$$\xi_t(i, j) = P(q_t = S_i, q_{t+1} = S_j|O, \lambda) \tag{2.37}$$

2.37 can be re-written in terms of the forward and the backward variables given in 2.22 and 2.26

$$\xi_t(i,j) = \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{P(O|\lambda)} \tag{2.38}$$

$$= \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)} \tag{2.39}$$

Define

$$\gamma_t(i) = \sum_{j=1}^{N}\xi_t(i,j) \tag{2.40}$$

Summation of $\gamma_t(i)$ over time index t can be interpreted as the expected number of times that state S_i is visited which is equivalently the expected number of transitions made from state S_i

$$\sum_{t=1}^{T-1}\gamma_t(i) : \text{expected number of transitions from } S_i \tag{2.41}$$

Similarly, if we sum $\xi_t(i,j)$ over time, we obtain the expected number of transitions from state S_i to state S_j as given in (2.42)

$$\sum_{t=1}^{T-1}\xi_t(i,j) : \text{expected number of transitions from } S_i \text{ to } S_j \tag{2.42}$$

With this definitions and observations the reestimation method for obtaining HMM parameters can be constructed as follows:

$$\overline{\pi}_i = \text{expected frequency(number of times) in state } S_i \text{ at time } (t=1) \tag{2.43}$$

$$= \gamma_1(i) \tag{2.44}$$

$$\overline{a}_{ij} = \frac{\text{expected number of transitions from } S_i \text{ to } S_j}{\text{expected number of transitions from } S_i} \tag{2.45}$$

$$= \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \tag{2.46}$$

$$\overline{b}_j(k) = \frac{\text{expected number of times in state } j \text{ and observing symbol } V_k}{\text{expected number of times in state } j} \tag{2.47}$$

$$= \frac{\sum_{\substack{t=1 \\ \text{s.t.} O_t = V_k}}^{T-1} \gamma_t(i)}{\sum_{t=1}^{T-1} \gamma_t(i)} \tag{2.48}$$

The objective function that we maximize is as follows:

$$Q(\lambda, \overline{\lambda}) = \sum_{Q} P(Q|O, \lambda) \log P(Q|O, \overline{\lambda}) \tag{2.49}$$

where Q is the state sequence.

The algorithm continues until finding the best sequence of $S_1 ... S_k$ and goes to model parameter estimation step.

It is known that Baum-Welch algorithm improves the value of the objective function at each step and converges to a local optimal.

CHAPTER 3

LITERATURE SURVEY

Human Action Recognition from visual sources is a challenging and complex problem for which there are several techniques proposed in the literature. The challenge of the problem does not only come from the unexpected human behavior, but also comes from the side effects caused by occlusion, moving background, clutter, illumination, non-rigidity and loss of 3D information.

In the literature, there are several approaches for human action recognition from visual resources. In this section, we mention about these approaches and make a comparison to our approach.

To begin with, action recognition problem from visual resources have the following components: feature extraction, action learning and classification, and action recognition and segmentation [52]. In this work we divided the problem into 3 major sub-problems which are feature extraction, tracking and recognition of action where recognition includes learning and action recognition phases.

This paradigm heavily relies on the performance of the previous process. In the feature extraction step, the discriminative and reproducible features have to be selected since they affects performance of tracking. Besides, the accuracy of tracking is very critical since the recognition step uses the outputs of the tracking phase which is not reliable in a cluttered environment. There are several expression of recognition methodologies using different criteria discussed in 3.1.

In literature review section, the literature survey for each sub-problem and comparison to our approach is given in detail.

3.1 Review on Human Action Recognition

Human action recognition is an active with a wide range of research made and a diverse field of real world applications such as surveillance, recognition of abnormal behavior etc. From computer vision point of view, action recognition is finding the best matched pattern, which is obtained previously in training phase, to the current observation. Here, the critical issue is the way of representing the pattern and the way/approach of recognition.

Action recognition from visual sources with different individuals brings difficulties due to variations in motion performance, recording settings and interpersonal difference [52]. Another difficulty comes from the fact that recognition is severely affected from the accuracy of tracking since it uses the tracking output directly for recognition purposes.

There are single layer and multiple layer(hierarchical) approaches to human action recognition problem [1]. Single layer approaches deal with single human body and gesture recognition whereas hierarchical approaches are generally applied to more complex problems such as human interactions and group activity cases. In this work, we are interested in recognition of single human body so make a review of single layer approaches. There are various single layer approaches one of which utilizes 3D cumulative of image sequences (called space-time) and the other uses the information between sequence of images (called sequential) [1]. Figure 3.1 show the decomposition of human recognition approaches and show the solution domain of the thesis. In this work we handle the recognition problem as a sequence of observations instead of 3D space time volume which corresponds to a sequential approach.

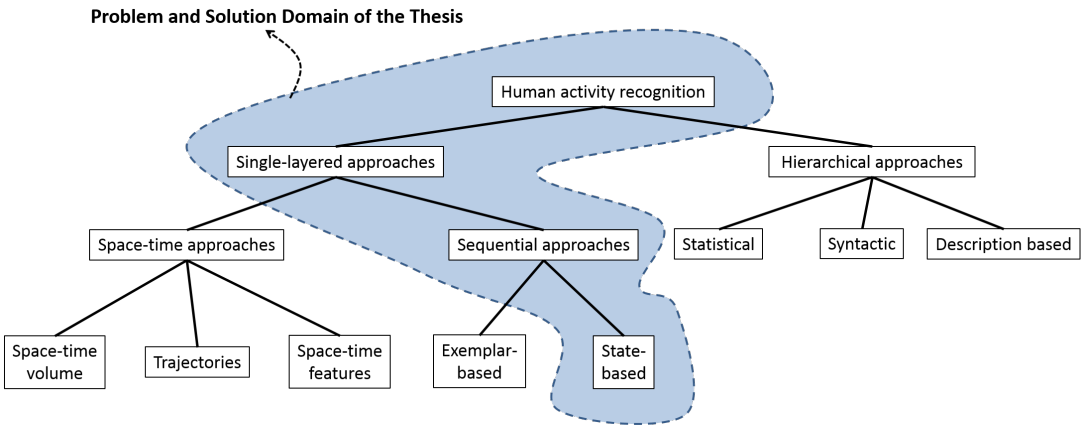


Figure 3.1: Problem and Solution Domain of HAR [1]

In this part, the literature survey for single layer approaches are presented. Considering 3D space-time approaches, the extracted features may contain history information. One of the approaches, [54] extracts tracklet descriptors. Firstly, the features are extracted and tracked by means of KLT Feature Tracker. Then the histogram of oriented gradients (HOG) and histogram of optical flow (HOF) properties of the tracked points on each trajectory are obtained and tracklet descriptors for each trajectory longer than 3 frames is constructed. The method utilized the bag of words technique and project the each tracklet to the closest dictionary element. Then the SVM classification is performed by histogram these words. The approach has 94.5% recognition performance with leave-one-out evaluation on KTH database. Note that, we utilize this method for benchmarking in occlusion scenarios. Another approach,[22], uses motion history image (MHI) and combined it with feature HoG which contains information on directions and magnitudes of edges and corners. The classification is performed by simulated annealing multiple instance learning support vector machine (SMILE-SVM) to acquire a global optimum. This approach gives a 100% success in CMU database.

Ziaeefar st al. [76] propose a Cumulative Skeletonized Image (CSI) which is constructed by combining the skeleton centers along time in a polar coordinate. The classification is performed by analyzing this angular/distance histograms and using hierarchical SVM.

Accumulated Motion Image (AMI) is proposed by [33] which is a spatio temporal feature composed of average of image differences. The distance between rank matrices of query and candidate videos are used for recognition purposes. AMI concept is motivated by Gait Energy Image [18]. There are efforts using Canonical Correlation Analysis (CCA) as a measure between videos with 95.33 % performance [32] and Principle Component Analysis (PCA) to one cycle of repetitive action in the video [37].

The approaches using space-time trajectories, one of which [68], extract dense interest points and track them by displacement of information and obtain dense trajectories of these points. They make use of HoG, Histogram of Optical Flow (HoF) and propose Motion Boundary Histogram (MBH) as local descriptors around the interest points as features. They make recognition by means of these features and obtain 94.2 % performance on KTH database. The method of [46] uses Harris3D interest points and extracts the feature trajectory using KLT tracker. The trajectories are represented as sequences of log-polar quantized velocities. The feature utilized in recognition is based on velocity history of tracked points and performs recognition by means of utilizing a generative weighted mixture model. This methods has a 74 % performance on KTH database.

The space-time local features are described as the descriptive points and their surroundings in 3D volumetric data with unique discriminative characteristics. They can be categorized as sparse and dense. The most popular sparse features are 3D Harris Detector [35] and Dollar detector [13] where Figure 3.2 shows the features of the latter which are obtained by Gaussian smoothing kernel and Gabor filtering. Methods using optical flow are examples of the dense case. Local features provides increased robustness to noise and pose variation and it compares the cuboid features for SVM and nearest neighborhood (NNC) classification methods obtaining over 80 % performance in [13].

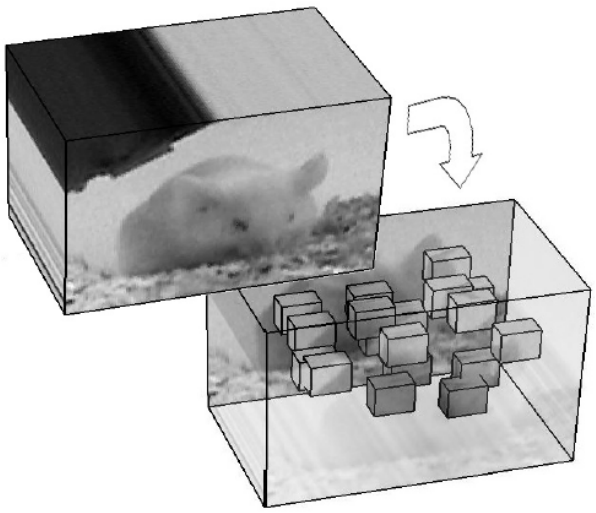


Figure 3.2: Cuboid Feautes of [13]

Studies in [28] and [6] improves performance of [13] by utilizing clouds of the proposed interest points. [28] uses k-means to cluster the clouds and perform classification by proposed Asymmetric Bagging and Random Subspace Support Vector Machine (ABRS-SVM) with a 95.3 % performance on KTH database. On the other hand [6] uses clouds which are obtained at different temporal scales given in Figure 3.3.

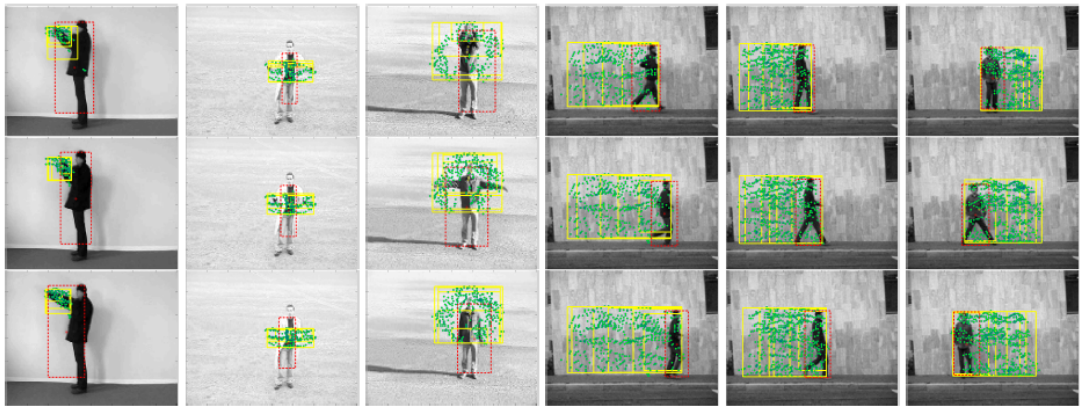


Figure 3.3: Examples of clouds of interest points at different scales [6]

The features proposed in [6] are given in Figure 3.4 with comparison to features of [13] which uses local information within a small region and sensitive to video noise and has low performance in zooming cases. The proposed features are extracted by first taking frame difference for understanding of focus region, then utilizes 2D Gabor filtering on the detected regions of different orientations which identifies salient features as given in Figure 3.4 and has 93.17 % performance on KTH videos.

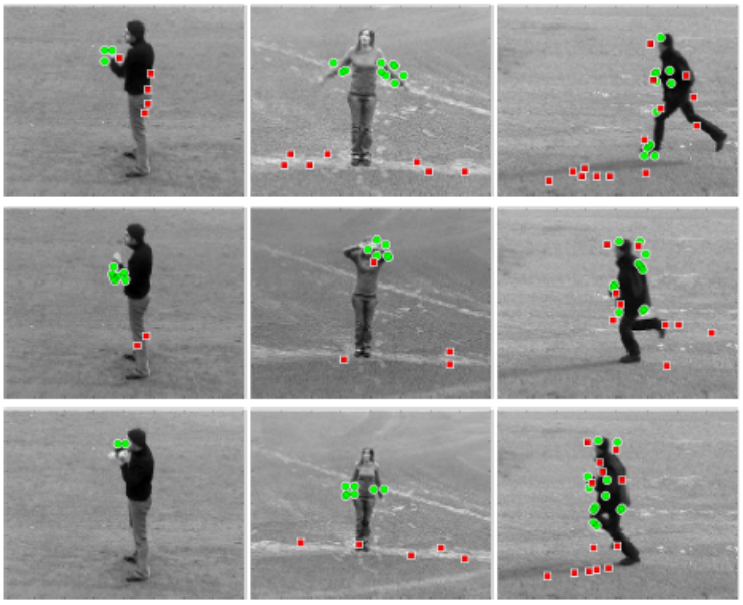


Figure 3.4: Visualization of cuboid based behavior recognition [6]

Harris3D [35] features are utilized in [63] in which classification is made by PCA-SVM and has 93.83 % performance on the KTH database. In order not to suffer from sparsity, [16] uses dense 2D Harris corners in multiple scales with two scale hierarchical grouping. In [58] 2D Harris Corners are extracted in each frame, as in our case, and local features are defined on log-polar histograms by using temporal similarities as given in Figure 3.5

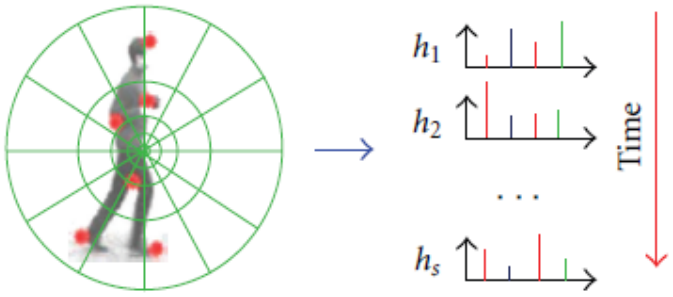


Figure 3.5: Visualization of Features of [58]

The flow chart of the approach [58] is given in Figure 3.6. It uses gravity center change property for classification and it is done by SVM and it has a 93.6 % performance on KTH database.

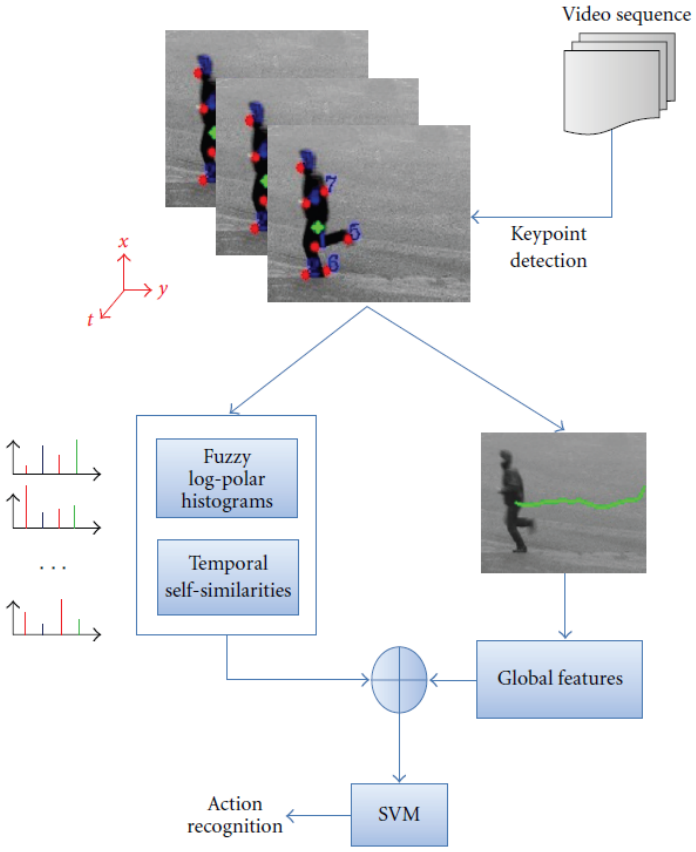


Figure 3.6: Flow-chart of [6]

The 3D local features are similar to the features used in this study in the sense that we extract 2D Harris corners in each frame which somehow corresponds to exploring 3D local features.

Optical Flow features are used in [24] combined with the shape feature. Location of interests are found by Multiple Instance Learning (MIL) framework which takes all feature channels as inputs. In our work we also combine velocity information along with the relative position information.

In [39] a completely different space time local feature set is used for recognition. [39] makes use of tangent bundle representation on a Grassmann manifold. They represent the video with third order tensors and performs classification by means of tangent vectors which are obtained by factorization of 3rd order tensors to a space of tangent spaces. They manually extracts human action from videos in

KTH database and re-size the videos to $20 \times 20 \times 32$ windows. Their approach uses leave-one-out cross validation technique and achieves a 97 % success of recognition in KTH database.

When we investigate the sequential approaches in Figure 3.1, in which the temporal relationships of observations are extracted and used in recognition phase, there are exemplar-based and state-model based approaches. In the exemplar-based approaches the representative template sequence of the video is used for recognition. In this approach, the difference in the velocity of performing the same action is handled with Dynamic Time Warping (DTW) as used in [12] and [65].

A shape-motion feature prototype is proposed in [36] as given in Figure 3.7. After shape-motion descriptors of interest regions are computed, action prototype is learned by k-means clustering.

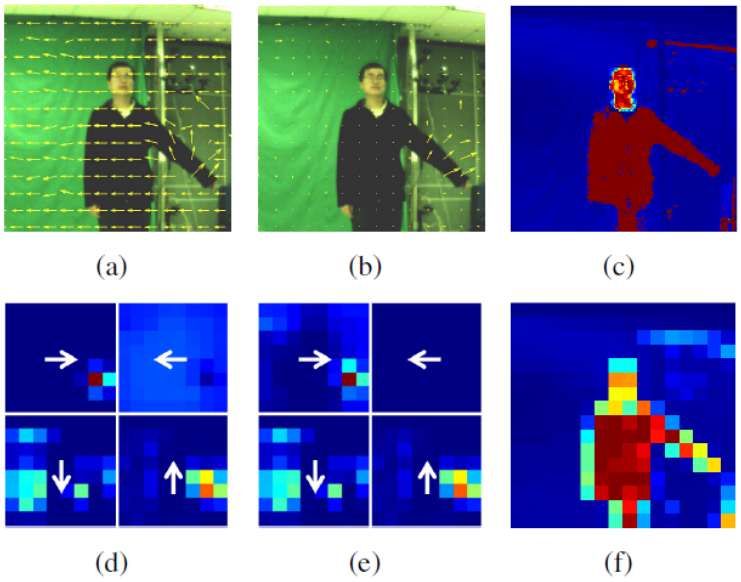
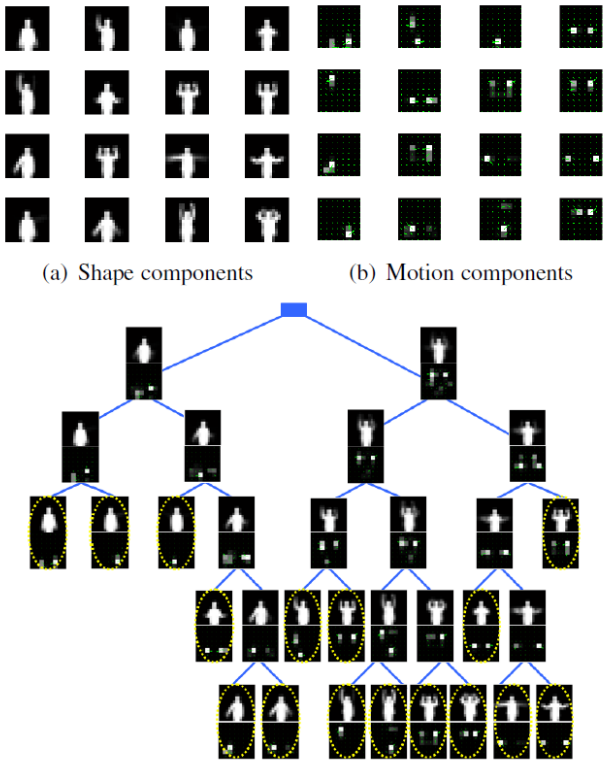


Figure 3.7: An example of computing the shape-motion descriptor of a gesture frame with a dynamic background. (a) Raw optical flow field, (b) Compensated optical flow field, (c) Combined, partbased appearance likelihood map, (d) Motion descriptor Dm computed from the raw optical flow field, (e) Motion descriptor Dm computed from the compensated optical flow field, (f) Shape descriptor Ds. [36]

Then a binary prototype tree is formed by a hierarchical k-means clustering using the prototypes as given in Figure 3.8. In testing phase, after detecting and tracking human body, prototype matching is performed by maximizing a joint likelihood of the actor location and action prototype. Note that the body center is determined by the height of the human bounding box, and side-length is proportional to the height of the bounding box and the shape feature is extracted from foreground which is obtained by different techniques in different backgrounds. Under static backgrounds binary silhouettes are obtained by background subtraction, appearance-based likelihoods or probabilities are utilized in dynamic one. The performance of the approach is 95.77 % in KTH database.



(a) Shape components (b) Motion components

(c) Binary prototype tree

Figure 3.8: An example of learning. (a)(b) Visualization of shape and motion components of learned prototypes for k = 16. (c) The learned binary prototype tree. Leaf nodes, represented as yellow ellipses, are prototypes. [36]

Talking about state model-based approaches, general trend is to represent each action by hidden states. In this area, utilization of HMM is very common as given in [5] and [72]. One of the drawbacks of HMM is modeling the duration of action and determining the transition matrix among this information.There are also extension of HMM techniques proposed as Coupled Hidden Semi-Markov Model (CHSMM) to overcome this problem and model duration of human activities as given in [40] and [48]. Transition of the synthetic poses is represented by a graph model called Action Net where each node contains a keypose which is the the 2D representation of one view of an example 3D pose as given in Figure 3.9.

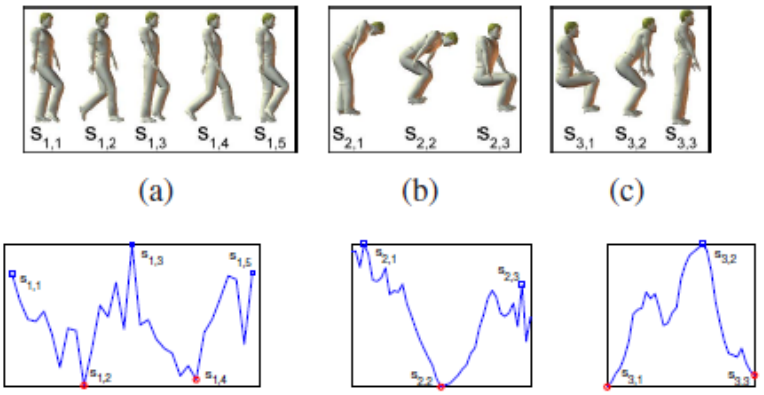


Figure 3.9: Automatically extracted key poses and the motion energy chart of three action sequences[40]

The HMM structure in [40] is given in Figure 3.10 where each node represents one key pose and an action composes of a chain of the extracted keyposes.

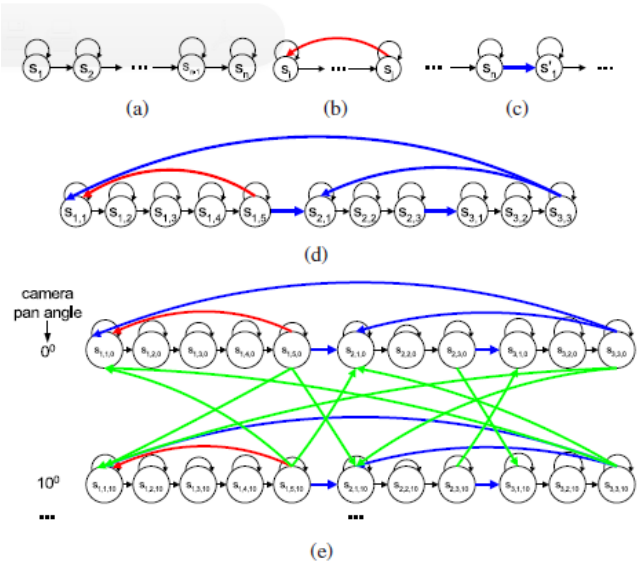


Figure 3.10: Action graph models. (a) The general model of a single action; (b) Back-link (in red); (c) Inter-link (in blue); (d) A simple Action Net consisting of the three actions. Each node contains a keypose which is the the 2D representation of one view of an example 3D pose as given in Figure 3.9 (e) The unrolled version of (d). Only models with the first two pan angles are shown.[40]

Posture representation is constructed by flexible star skeleton and human extremities are matched using contours and histograms of the posture in [73]. The polar segmentation used to extract features and to obtain a histogram, is given in Figure 3.11. The action is also recognized by HMM which has high performance on Human Climbing Fences dataset.

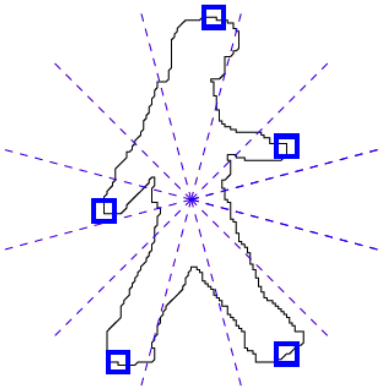


Figure 3.11: A simple histogram to extract feature vectors from frames.[73]

Human movements are described with dynamic texture features in [29] and recognized by HMM which has 93.6 % performance in KTH database. In [61], a discriminative Semi-Markov model approach is proposed for action recognition which has 95 % performance in KTH database.

Comparing the existing literature with our approach from the object representation and recognition points of view, we can state the similarities as we use Harris Corner Detector and utilize both velocity and relative position as features. Furthermore we used HMM for human action recognition which have significant performance results. But there is no approach performing group tracking using GM-PHD filter and making recognition with state estimates of GM-PHD and GM-PHD itself.

3.2 Review on Hidden Markov Models

HMM is a suitable tool for modeling the variations in the observations of an action and for discriminating among different actions [48]. There are several studies in which HMM and its extensions are used for recognition purposes not only for human action but also in other areas.

There are some drawbacks of HMM one of which is modeling the duration of

time. Besides decision of number of hidden states and observation symbols is very problem dependent. When we investigate the learning problem, it finds the local maxima which is one of local maxima of a complex domain.

HMMs are basically a class of Dynamic Bayesian Networks (DBN) where there is a temporal evolution of nodes. The elements of HMM and detailed background information is given in Section 2.3. Hierarchical HMM (HHMM) is proposed for the gap between low level data and high level semantics which utilizes HMM for action recognition which models complex multi-scale structure in [8]. Abstract HMM (AHMM) where each state is dependent to a hierarchy of actions is described in [9]. In order to overcome the duration problem, Hidden Semi Markov Model (HSMM) which has explicit state duration models is introduced and is applied to video events in [21]. Switching Semi Markov Model (S-SHMM) is proposed in [14] which is a two layer extension of HSMM and applied to recognition.

There are many approaches proposed to improve the performance of HMM. However, in this work we utilize standard HMM since we use quantized values of features as observation symbols and has constant frame size for each group.

Focusing on the representation of observations shows that there many different code-book generation mechanisms. In [72], the code-book is generated by selecting 11 number of representative frames of the whole action. The ratio of white pixels in 8×8 regions are calculated and the sequence of these features is used as the feature in symbol construction.

[66] utilizes 17 dimensional feature set for recognition purposes using HMM with 3 hidden states and 5 Gaussian Mixtures for each observation symbol. It also extracts projection histogram features of vertical and horizontal axis which is said to be sufficient in order to infer posture of the person and make a comparison of the two feature sets in two databases. 17 dimensional feature set has higher performance on Weizmann Data where the latter feature set is better in UT-Tower data set.

The affect of code-book size, clustering methods (K-means vs LBG) and feature

selection on recognition performance is investigated in [64]. It is concluded that the best performance is obtained with code book size greater than 30 in LGB clustering with IC-based shape features using LDA. Note that in our case, we selected the code book size as 11 which correspond to mapping all features into the 0-10 interval.

CHAPTER 4

TRACKING

Tracking is simply a process which generates the trajectory of the object being tracked. There are different approaches to track human bodies in the video. Tracking can be performed by detecting the possible objects in each frame and establishing correspondence between the detected objects in the consecutive frames. Another approach is to make detection first, then updating the track by estimating the new state of the object(its extracted parameters) in the following frame iteratively.

The object states and the motion capabilities of the human are determined by the object representation approach, so the tracking method is highly correlated to the object representations type.

The tracking methods can be categorized into two: *probabilistic* and *deterministic*.In this thesis, we utilize a probabilistic technique for group tracking called GMPHD which both group track the body as a complete object and constructs state estimates of individual features/measurements for tracking purposes. So an intensity map is obtained by a combination of Gaussian Mixtures belonging to the measurements/features.

The solution domain utilized in this work is given in Figure 4.1.

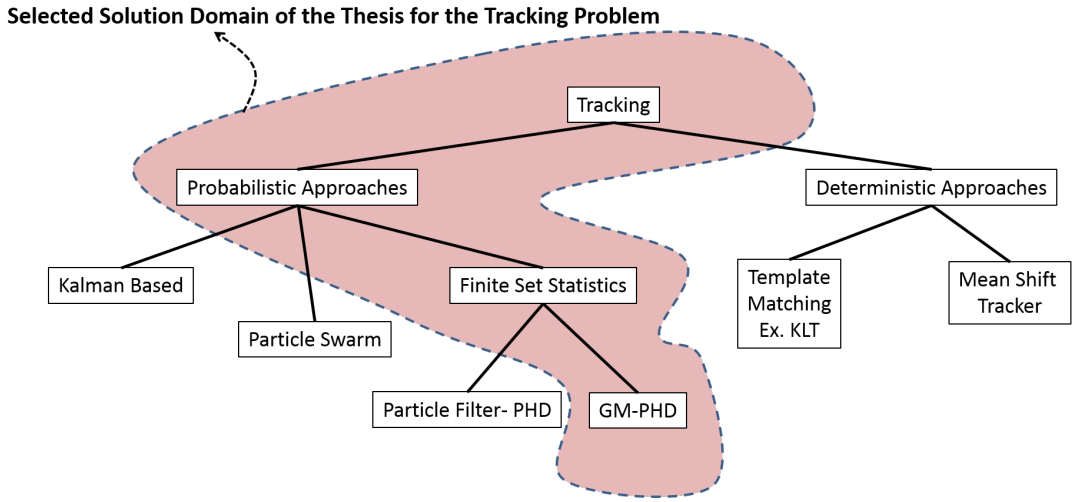


Figure 4.1: Selected Solution Domain of the Thesis for the Tracking Problem

4.1 Probabilistic Approaches

There is unpreventable noise existence in visual tracking. The source of the noise might be the visual sensor as well as the object behavior itself. In order to take this in-suppressible noise into account, the probabilistic approaches are utilized and some different estimation techniques are used. In the case of multiple target existence, in addition to these estimation techniques, different association techniques are proposed. In this part, brief information on these techniques is provided. In probabilistic approaches, the most frequently chosen object representation type for object tracking is the point representation. The state of the object is often chosen as the location and velocity of the object and tracking becomes the operation of finding the most probable state in the next frame (prediction) and update this prediction by the new observation (correction). In probabilistic approaches, as well as point representations, the shape and counter information/representation is also common.

Probabilistic approaches can be categorized into three classes depending on how the problem is modeled:

(i) Classical Probabilistic Approaches : Kalman based

(ii) Particle Swarm Optimization Techniques with Particle Filter

(iii) Random Set Statistics : Finite Set Statistics

4.1.1 Classical Probabilistic Approaches

To begin with the single target case, the most common approach is Kalman Filter and its various versions. Briefly speaking, Kalman Filter is a linear estimation technique where the state variables are assumed to be Gaussian distributed. The first step is the prediction step in which the prediction of the current state is found from the target dynamics and previous information. Then as the new measurement comes, the prediction is updated using the correction step. These prediction and correction steps are coupled and performed iteratively for each new incoming data. Kalman filter has been a quite popular technique in vision for a very long time [7].

Particle Filter Method is proposed to overcome the Gaussianality and the linearity constraints of Kalman filtering [34]. In particle filter, the distribution is approximated by the samples and these samples are taken into account instead of the whole distribution. This approach breaks the Gaussian probability distribution constraint of KF. The samples known as particles are Dirac-delta functions and are weighted (with observation), predicted and corrected which is similar to classical Kalman filter in fitness.

Another classical statistical tracking approach is using some data association techniques. Probabilistic Data Association (PDA) technique takes each potential target as a track candidate and the statistical distribution of the track error and clutter is used in this algorithm. This method considers only one of the measurements as a track where all the others are taken as false alarms.

Considering Multiple target scenarios, there is a necessity to perform an association mechanism between the tracked objects and the measurements. After extracting the association information, the single target approaches are safely applied to each target. There are many association mechanisms for multiple target case. For instance, Joint Probabilistic Data Association (JPDA) is an

extension of PDA that is capable of representing multiple targets by associating all measurements with each track. [55] uses a version of this method for region tracking. The disadvantage of this method is that it is not able to track when the number of tracked objects changes in the scene.

Similarly, [47] models data association and probability of target/track presence with a recursive probabilistic approach called integrated probabilistic data association (IPDA) which is said to have less computational complexity with almost the same tracking performance with PDA [31]. Multiple Hypothesis Tracking (MHT), proposed by [56], is another association algorithm which is capable of handling the change in the number of objects in the scene. MHT considers all possible association hypothesis so in a few iterations the number of possible tracks become computationally untractable. There is another probabilistic approach, called Probabilistic Multiple Hypothesis Tracking (PMHT), proposed to reduce the computational load of the algorithm by taking the associations as statistically independent random variables in [62]. [23] proposes a method for multiple tracking that performs state estimation by particle filter and association by a method similar to PMHT. [10] performs whole human body tracking with this multiple hypothesis point of view.

[26] proposes a different approach which considers tracking with particle filter using 3D cylinder object representation. In this method, both the background and foreground are modeled as mixtures of Gaussians, and tracking is performed by particle filters whose parameters are the shape, velocity and 3D positions of all the objects in the scene. Even if the maximum number of object in the scene is defined, the method is able to handle birth and death of target as well as occlusion. The drawbacks of the method are those it uses the same template (cylinder) for each object, and requires training for each object. The silhouette object representation for the object to be tracked gives the advantage of employing complete and large variety of object shapes. There are two main schemes for silhouette representation. The first one utilizes the whole contour of object region and assigns binary indicator to this region. Contour can be defined by some control points or by means of a function defined on a grid.

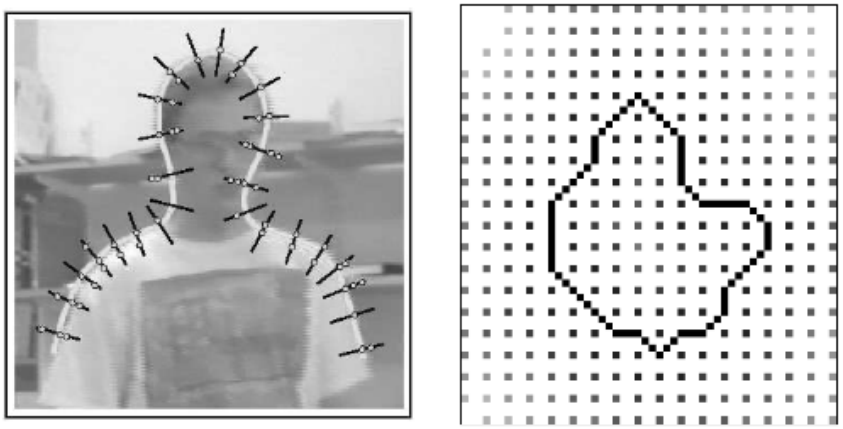


Figure 4.2: Explicit/Contour and Implicit/ Grid Representations of Silhouette [42]

[25] uses contour of the object for tracking. It models the contour shape and motion as a state vector. This well-known condensation algorithm tracks the object by means of particle filter method and the parameters of the particle filter are determined by the contour of the object which is a new approach to tracking scheme. [41] extends this approach to multiple objects by adding an occlusion handling mechanism. Figure 4.4 shows some results of the condensation algorithm.

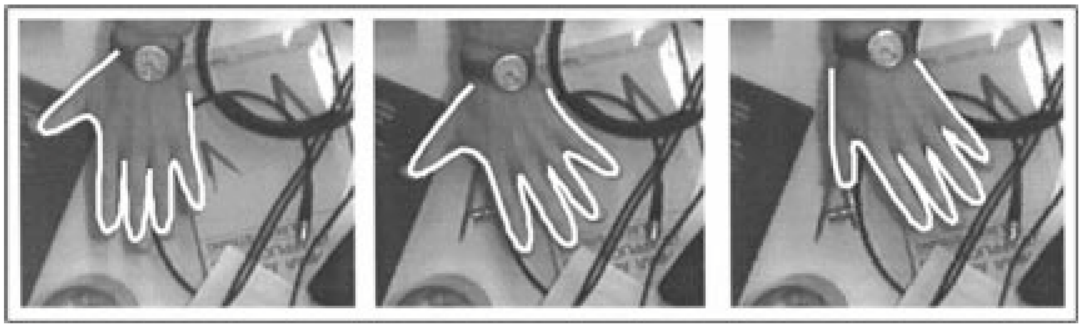


Figure 4.3: Some Results of Condensation Approach, [25]

[31] brings a new approach for multi-target tracking capabilities of particle filtering methods. When the number of target high, the traditional joint particle filter is unusable due to low track quality and high error. An MCMC-based particle filter is proposed which brings an improvement on particle filter that makes it capable of handling interacting targets. The MCMC-based particle filter increases the tracking performance compared to standard PF and decreases the failures reported with the advantage of requiring less number of samples to track the joint target state. On the other hand this method fails when the targets in the scene overlap which is a considerably common situation in human body tracking when the whole human body parts are taken as objects to be tracked.

There are tracking algorithms specifically focused on non-rigid object. The tracking of a non-rigid object is performed mainly with energy function constraint or by using Kalman filtering under the assumption of Gaussian noise and knowledge of dynamical model of the object.

In case of deformable non-rigid objects, the object model and target distribution are nonlinear and non-Gaussian. So, simple Kalman based tracking assumptions may not be performed. To track non-rigid objects, [50] proposes a vision system that is able to learn, detect and track the features using mixture particle filters and AdaBoost. This article shows the performance of the proposed algorithm from a hockey match where various number of hockey players in the scene. In this method, a mixture particle filter is used for every object being tracked and

HSV color space is used to make the tracking more insensitive to illumination effects, since the intensity value is reasonably independent of color. The detection is performed by Bayesian multi-Blob detector and tracker is fed by the updated detection models. To increase object modeling approximation quality, the objects are divided into sub-regions according to color and spatial location and color histograms are created for each sub-regions to form the observation model.

[57] introduces Kalman and particle filter usage for lower human body part tracking considering the human biometrics. The tracking is performed with 2D model constrained by the human bipedal motion. The experiments performed in indoor and outdoor sequences give promising tracking results.

4.1.2 Particle Swarm Optimization Techniques with Particle Filter

Particle Swarm Optimization is a population based heuristic optimization technique which iteratively tries to improve a candidate solution with respect to a given constraint/objective function [30]. It is affected by the organisms behavior/attitude in which each organism make individual motion and able to move together successfully as a group. The candidate solutions are the particles and this technique try to update each particle state (position, velocity) by considering both best local and global positions. In this way, while each particle moves to their best position, the swarm of particles does not get separated from each other with the update by the influence of global best position. Figure 4.4 shows a result on global and local best points.

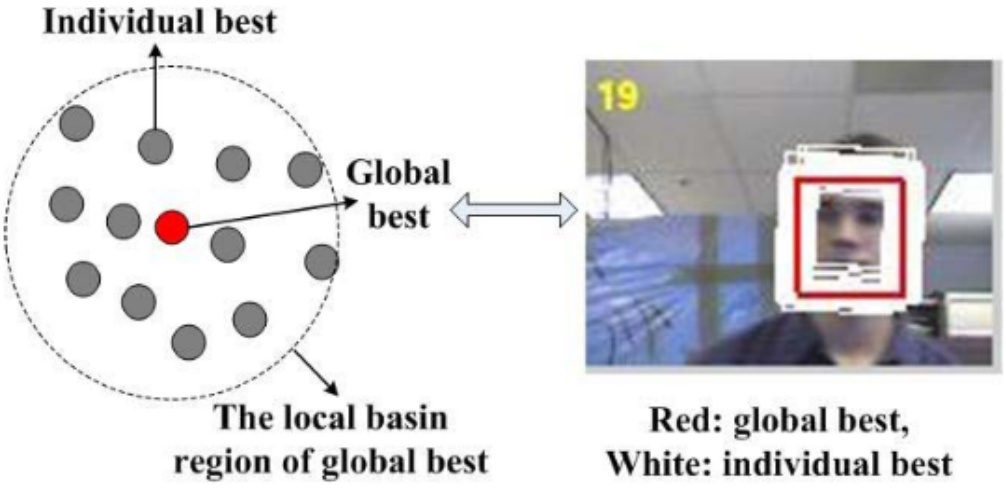


Figure 4.4: Convergence Criteria for Sequential Particle Swarm Optimization, [3]

There are several papers proposing the utilization of PSO approach for human tracking in a probabilistic framework. For instance, [69] proposes PSO based particle filter approach for human head/face tracking. One of the problems of PF is the particle improvishment problem which arises from uniform re-sampling in PF and mostly solved by increasing the number of particles extensively which increases the computational cost. This work utilizes mutation operator and combine it with PSO which helps redistribute particle to their close local modes of the posterior and increase the diversity of particle which helps overcome the improvishment problem of PF. By this approach, a more robust and low computational cost algorithm is proposed for single target tracking.

[3] introduces a prey-predator scheme for human tracking. In this approach, the prey pixels form a kind of template and the a swarm of predator particles track the scent (color) of the prey particles by individual and group behavior rules. Particle movement is defined by a combination of color, topography, swarm centroid, swarm velocity, particle velocity and swarm centroid prediction properties. The tracking performance is determined by adjusting the weights of these properties. This work only handles single object tracing problem and does not deal with multi-target tracking scheme.

As [3], [74] uses swarm intelligence perspective for single object human and face tracking.

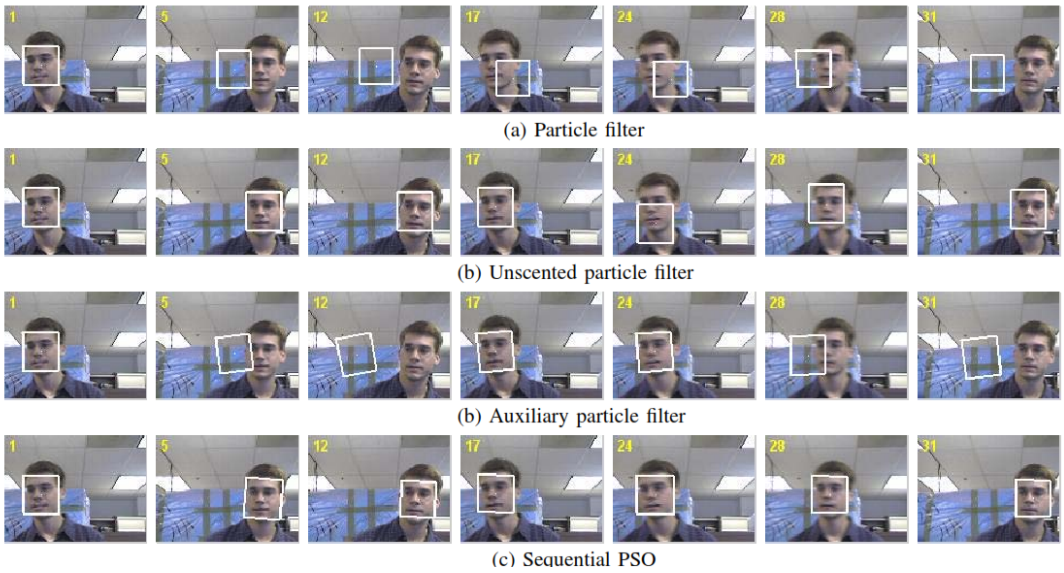


Figure 4.5: Comparisons between variants of PF and sequential PSO, [3]

Figure 4.5. shows a comparison between PF based methods and PSO approach for face tracking.

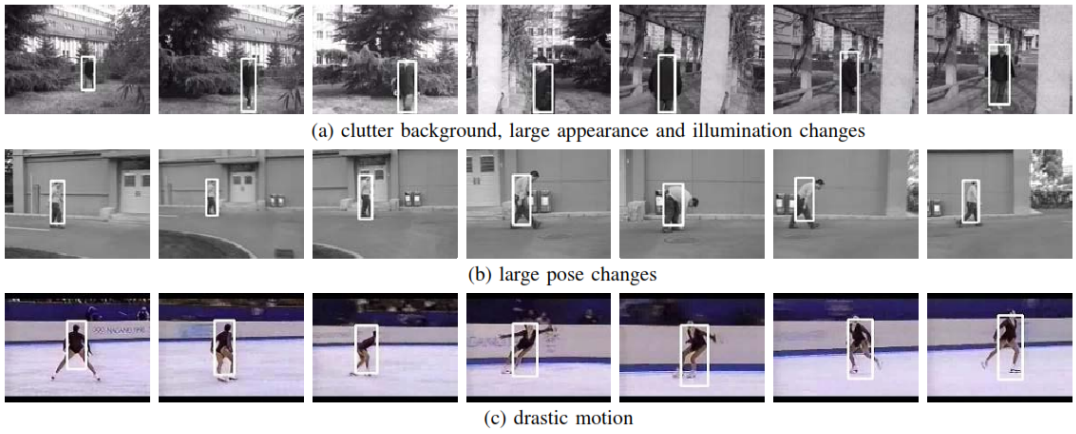


Figure 4.6: Experimental Results on Single Object Human Tracking, [74]

Figure 4.6 shows the capability of PSO in human tracking. It is easily seen that PSO is able to handle clutter, pose and silhouette change and sudden movements of human body for single object.

Unlike [3] and [74], [15] compares PSO, classical Particle Filter and Enhanced Particle Filter methods in human tracking and propose that EPF tracker has a better performance in term of occlusion handling and computational cost.

There are also methods proposed using PSO for multi-target tracking problem [75, 49].

The advantage of PSO is that it uses the the prior information that the particles move in coordination unlike utilization of independent particles in the conventional particle filter method. So PSO is a very meaningful approach in human tracking since the human body parts move as an inseparable parts of the whole body.

4.1.3 Finite Set Statistics

Finite set statistics (FISST) is a novel and the first systematic approach for multi-sensor, multi-target tracking [17, 44]. The major problem of FISST was its combinational complexity. This problem is solved by the Probability Hypothesis

Density filter approximation proposed by Mahler [45].

[71] proposes to use Particle Filter implementation of the PHD filter for multi-target tracking. The reason of using particle filter implementation of PHD is the use of nonlinear human model. The approach first extracts foreground objects by background subtraction method. It then takes the centroids of the foreground objects as measurements for PHD filter. The algorithm results show that this approach is successful in handling new human entrance, exit and occlusion. The background in the videos is relatively complex, and birth, merging and disappear situations are handled successfully. Although the background is complex, it is fixed in all the frames which make foreground object extraction highly easy and increase tracking capability.

[43] also utilizes particle based PHD Filters for multi-target tracking. Then applies K-means method is used for particle clustering, and using the centroid information of these clusters, association is performed by graph matching to improve the robustness of the tracking performance.

[70] uses Gaussian Mixture-PHD filter method for multiple human tracking. The background is modeled statistically and the background characteristics is constructed by the employment of spectral, spatial and temporal features. After modeling the background, foreground objects are extracted and used as measurement for GM-PHD filter. The advantage of this method is that it models birth, survival and death process by number of Gaussian component information. This paper uses the same CAVIAR datasets as in [71] for algorithm testing. This method GM implementation instead of particle filter approach, so the computational complexity of the algorithm is much [51] employs GM-PHD filter for tracking for human tracking from aerial videos. This paper uses aerial videos, so it deals with homographic correction of the scenes before detection. It uses RANSAC to handle outliers and performs homography transformation to deal with camera motion. After handling camera motion, the detection is performed by intensity-based criteria and GM-PHD filter is used to track the detected objects. This method also utilizes Gaussian Mixture Cardinalized PHD filter which is a kind of simplified second order moment tracking that propagates car-

dinality information. [20] represents the human bodies with moving blobs and track these by means of PHD filtering approaches and estimate the unknown number of persons in the frame. The major advantage of PHD approach is that it handles association of objects by considering the effect of the newly coming measurement to all tracks. This method proposes two different smoothing technique to PHD filtering to overcome the miss detection bias and occlusion.

4.2 Deterministic Tracking

Deterministic tracking approaches is mainly grouped into two and detailed information for these techniques will be provided in this section:

(i) Feature/Template Matching Techniques

(ii) Mean Shift Tracker

4.2.1 Feature/Template Matching Techniques

Template based approaches are those in which the tracking is performed by measuring the similarities between the template and the current measurement. In these approaches the templates can be constructed in two different ways. One of them is to use pre-defined templates. Second one is to construct templates from video and dynamically update them with new information.

The template matching technique is simply a searching mechanism in the image for a region similar to the object template which is defined beforehand. There exist some similarity measures such as proximity, cross correlation, color/intensity difference. The main aim of template matching is to find the location and occurrence of the object. The critical point here is in determining the invariant feature/property of the object.

The representation of the template may alter due to object representation. In case of patch object representation, the property of the region inside the

patch(rectangular, circular) is used and similar region is searched in terms of the selected property.

There are various versions of patch matching one of which categorize features as stable, transient and noise [27]. In terms of face tracking of a speaking person, the stable features are the stationary features as representing nose, transient ones are the moving features around mouth. The outliers are assumed to be noise. These features inside the patch are used to define the translation and warping parameters of the object as well as tracking.

Another method uses Kanade-Lucas-Tomasi (KLT) tracker to find the translation of the region whose center is the interested point. The quality of the tracked region is determined by affine transformation output of the patch between consecutive frames.

In multiple object tracking with template matching, [26] proposes a multiple-layer approach in which different layers are constructed for background and each object(foreground) in the scene where each object layer comprises shape, intensity and motion models. Then each pixels probability of belonging to one of the layers is calculated by means of objects shape and motion characteristic and the layers are updated with new incoming image properties.

Although these methods are successful from single-view case, they fail as the object pose/view changes. There are two main approaches dealing with multi-view case. The first one is [4] which performs tracking of articulated objects by eigen-tracking. In this method object appearance model is constructed using Principle Component Analysis (PCA) and the image is transformed to the eigen-space. The affine parameters which are the parameters of transformation from the current object image to the image reconstructed by eigenvectors are found. Other technique uses, [42], Support Vector Machine Classifier for tracking. In case the complete human shape is utilized in tracking, silhouette of the object is used. There are both deterministic and probabilistic approaches employing the silhouette of the object for tracking means. Briefly speaking about the deterministic approaches, [41, 2, 51] use shape matching schemes whereas [70, 20, 30] employ gradient descent techniques.

Tracking using KLT

Kanade-Lucas-Tomasi (KLT) Tracker is an image based tracking algorithm based on Optical Flow information in the image ([38, 39, 40]). KLT calculates optical flow using least squares criterion and assumes that the flow is constant around the predefined neighborhood of the interested pixel. The calculated flow is assumed to be hold for all pixels within a window centered at the feature point.

KLT tracker searches the best match of the interested pixel in the next frame. In general, the pixels are the features extracted by Shi and Tomasi's minimum eigenvalue method which is optimized feature extraction method for the tracker. The extracted features are fed to the tracker only in the very first frame and KLT tracks the interested point in the next frames. The features used by KLT are located by examining the minimum eigenvalue of each 2 by 2 gradient matrix, and features are tracked by minimizing the difference between the two windows.

KLT tracker uses optical flow information with some basic constraints. The basic assumptions that;

1. The flow is constant within a local neighborhood of the feature point of interest
2. The displacement between consecutive frames is small.
3. The intensity of the patch does not change between two nearby instants.

Comparing (group) tracking methodologies, the working principle of PHD and KLT is quite different. In KLT the features are extracted in the very first frame and tracked in the next frames by intensity based (in this case optical flow) techniques. So when KLT loses the points it is not possible to re-track of the points. On the other hand, PHD needs measurements in every frame. So, although the measurements are lost, it is possible to re-track the point if the feature is extracted in the next frames. In section 6.4, tracking capability of KLT and the proposed algorithm (GM-PHD) is compared in terms of recognition performance.The comparison is performed by changing only the tracking block of the proposed recognition approach. The feature extraction and recognition

methods are kept as the same. The fundamental structures of the algorithms are given in Figure 4.7.

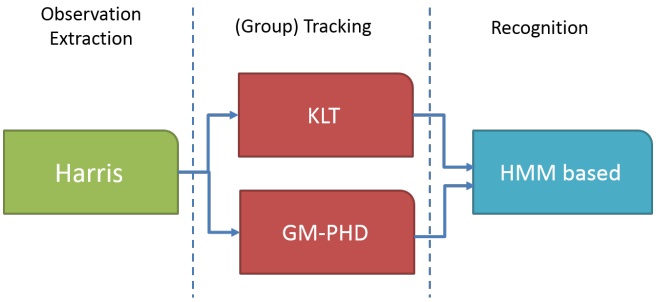


Figure 4.7: Structure of applied algorithms for comparison between KLT and GMPHD

The details of the experiments and the results are given in Section 6.4

4.2.2 Mean Shift Tracker

Mean-shift tracking algorithm employs the color/intensity histogram of the object being tracked. It compares the histogram of object in the current frame and histogram of candidate regions in the next frame and find the location of the object in the nexy image iteratively maximizing the correlation between two histograms. Object tracking for an image frame is performed by a combination of histogram extraction, weight computation and derivation of new location.

[11] uses circular patch for object representation and uses weighted histogram property. The method finds the location of the object by mean-shift tracker in which the similarity between the color histograms is maximized in terms of Bhattacharya coefficient. The advantage of this method is in the searching mechanism which decreases the computational complexity. Note that for an accurate mean-shift tracker the parameters/features should be fed accurately to the approach. So the initial selection/detection of the patch region has to represent the object perfectly. Figure 4.8 shows a result of a mean-shift tracker based approach. As seen it can track the shirt of the child on the image successfully using color information.

Figure 4.8: A Mean Shift Tracker Result, [2]

CHAPTER 5

PROPOSED ACTION RECOGNITION APPROACH USING GM-PHD AND HMM

Human Action Recognition from visual sources is a quite common approach which brings some difficulties from its nature. The appearance of the human in the video is severely affected by various factors like complete body motion, human pose change, independent human body parts' motion and the changes in the scene itself. The difficulty arises not only from these different motion affects but also the change rate of them. The human body is capable of making sudden movements of their bodies/pose as well as the individual parts of the body. In this work we recognize human body by modeling the global and local body motions.

Considering the video frames as our source, there happens to be information loss due to projection of 3D nature of human to 2D image view. So, although the motions of human body are linear, this projection makes the body motion nonlinear. In addition to this information loss, considering the fact that the human body is a non-rigid object, the form of the tracked object changes rapidly as the pose of the body changes. There are various different poses that the human body may take so the number of possible forms of the silhouette may be too extensive. As a trivial result, this makes the representation of the human body more difficult.

In addition to these, it is likely to face the occlusion problem of the human body. The occlusion may arise not only from any other object in the scene but also the local body parts' individual motion. Both of these occlusion types may severely

affect the appearance of the human body and should be taken into account in the model.

Human Action Recognition covers the detection, tracking and action recognition problem. In this thesis we propose a different solution to this problem utilizing GM-PHD filter which group tracks the measurements/observations of the human body as it evolves over time from probabilistic framework. For recognition, sequential information on human is gathered from visual sources and the best possible estimate of the hidden variables are investigated using HMM.

In addition to utilize classical state estimates of GM-PHD, GM-PHD distance and OSPA metric is taken into account for recognition purposes.

The flowchart of the proposed approach is given in 1.1 and 1.2.

In order to crystal clear the feature concept, I want to make some explanations. The image features and the ones used in recognition phase are different. The image features extracted using Harris Corner Detector are observations and utilized as measurement for GMPHD Filter. On the other hand, GMPHD filter estimates and their moments are taken as features for recognition and an unique HMM is constructed for each feature. The image features are denoted as observations.

5.1 Observation Extraction by Harris Corner Detector

Selection of observation extraction method is a critical and fundamental step in action recognition problem. This is because the more robust, discriminative and informative our observations, the higher performance of action recognition we obtain.

Observation extraction is the basic step of our action recognition problem. Since whole frame contains redundant information, we have to eliminate this redundancy and extract the discriminative information on human in action. The observation extraction method has to be chosen such that it is effected from video problems in minimum order. Considering the general problems, there are

occlusion, background change and illumination problems and loss of 3D information in video processing approaches. When we focus on our data set, there is a human body in action and almost constant background which have the following properties:

1. Articulated object
2. Occlusion of limbs
3. Zooming
4. Illumination change
5. Clutter
6. Shadow
7. Different human characteristics (carrying a bag, wearing a coat, male, female, with long hair etc.)

In this thesis, we choose Harris Corner Detector (HCD), [19], as the observation extractor which generates a point for which there are two different edge directions in a local neighborhood of the point. The advantage of HCD for articulated human body is that articulated parts are very likely to have corner points. So, we have consistent points extracted in consecutive frames. Besides, it is less affected from zooming, illumination change than the other extractor methods. On the other hand, it unfortunately extracts the observations belonging to the shadows and different human wearings which have to be eliminated during the recognition process.

The typical parameters selected for the proposed solution is provided in Table 5.1.

5.2 Group Tracking by GM-PHD

In the tracking phase, all HCD observations are fed to the GM-PHD filter as measurements for each frame. Then, GM-PHD filter produces target intensity

Table 5.1: HCD Parameters Selected for the Algorithm.

Parameter	Value
Standard deviation of smoothing Gaussian	1
Threshold	20
Radius of region considered in non-maximal suppression	1

which is a Gaussian Mixture function. Training and testing phases of action recognition is performed by extracting information from the Gaussian Mixture in time.

In this work, target and measurement models together with the related GM-PHD parameters are provided:

1. **State Model:** State vector is defined as $\psi = [x\ y\ v_x\ v_y]'$ and constant velocity motion model is selected to be as in (5.1).

$$\psi_k = F\psi_{k-1} + Gw_{k-1} \tag{5.1}$$

where

$$F = \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{5.2}$$

$$G = \begin{bmatrix} \frac{\Delta t^2}{2} & 0 \\ 0 & \frac{\Delta t^2}{2} \\ \Delta t & 0 \\ 0 & \Delta t \end{bmatrix} \tag{5.3}$$

$\Delta t = 1/25$ sec. and $w_k \sim \mathcal{N}(.; 0, 1)$.

2. **Measurement Model:** The measurement model is chosen to be as in (5.4).

$$z_k = H\psi_k + \eta_k \tag{5.4}$$

where,

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \tag{5.5}$$

and $\eta_k \sim \mathcal{N}(.; 0, R_k)$ with $R_k = diag([1, 1])$.

3. **Clutter (Background Noise) Model:** Clutter random finite set is assumed to be uniformly Poisson distributed as given in (5.6) over the scene area of 160pixel$\times$120pixel, with $\lambda_c = 5.2\times10^{-5}(\text{pixel})^{-2}$. This corresponds to an average of 1 clutter measurement over the scene of interest.

$$\kappa_k(z) = \lambda_c V u(z) \tag{5.6}$$

where $u(.)$ is the uniform density over the scene, V is the scene area, and λ_c is the average number of clutter return per unit area.

4. **Detection and Survival Probabilities:** Detection and survival probabilities, $p_{D,k}$ and $p_{S,k}$, are taken as same, which is 0.98.

5. **Birth Model:** No spawning is assumed and the spontaneous birth intensity is assumed to consist of Gaussian functions uniformly placed over the scene. Their means are selected to be placed at 10 pixels away from each other.

$$\gamma_\psi(k) = \sum_{i=1}^{16}\sum_{j=1}^{12} w_\gamma \mathcal{N}(x; m_\gamma^{i,j}, P_\gamma) \tag{5.7}$$

where

$$w_\gamma = 0.01/(12 \times 16) \tag{5.8}$$

$$m_\gamma^{(i),(j)} = [5 + (i-1) * 10, 5 + (j-1) * 10, 0, 0]' \tag{5.9}$$

$$P_\gamma = diag(5, 5, 2, 2) \tag{5.10}$$

where $i \in 1, 2, \ldots, 16$ and $j \in 1, 2, \ldots, 12$

Note that for the rest of the section, GM-PHD filter output at time k, $D_x(x,k)$, is assumed to be as in (5.11).

$$D_\psi(\psi,k) = \sum_{i=1}^{N(k)} w_i(k)\mathcal{N}(\psi; m_i(k), P_i(k)) \tag{5.11}$$

5.2.1 Utilizing Image Intensity Difference in GM-PHD

One of the novelties of this thesis is the use of the intensity difference of image blocks in GM-PHD group tracking phase. The intensity information in the image is not utilized in standard GM-PHD group tracking approach. There is visual and powerful information in image which has the capability of increasing the tracking performance of GM-PHD. So, we use intensity difference of the image blocks centered at GM-PHD states.

Considering GM-PHD as an intensity map, the weights of Gaussian mixtures are calculated as in Table A.2 of Section A. The first two elements of the state vector corresponds to the positions of the objects. This part of the state vector is called "position state". We assume that there is small illumination change around the Position state in-between the consecutive frames.

We use $m \times n$ blocks around the position state and compute the intensity difference between the image blocks. Then we calculate the mean square difference between the patches as given in (5.12).

$$\Delta(k) = \frac{1}{N_{nor} p_s} \left\| \Gamma_{k|k-1} - \Gamma_{k-1|k-1} \right\|_{\mathrm{F}} \tag{5.12}$$

where Γ is 2D image intensity matrix, N_{nor} is the normalizing factor and p_s is the patch size. F denotes Frobenius norm and the normalizing factor is selected as 51 assuming the maximum intensity difference between the pixels in consecutive frames is 255.

$\Gamma_{k-1|k-1}$ is chosen as a $5x5$ intensity matrix of the gray scale image. The position of the center element $\Gamma(3,3)$ is located at the position estimate belonging to time

instant at time k, i.e. $\widehat{x}^{P}_{k-1|k-1}$ and the center element of $\Gamma_{k|k-1}$ is chosen at the predicted state estimate, i.e., $\widehat{x}^{P}_{k|k-1}$.

The weights in the 'Measurement Update stage' of the GM-PHD flow is calculated by including the intensity difference, $\Delta(k)$, into the related equation, (5.13).

$$w_k^{(lJ_{k|k-1}+j)} = p_{D,k} w_{k|k-1}^{(j)} \mathcal{N}(z; \eta_{k|k-1}^{(i)}, S_k^{(j)})(1/\Delta(k)) \tag{5.13}$$

$\Delta(k)$ term is included in the equation (5.13) as division because of intensity difference is inversely proportional to weight. If the intensity difference between the patches is high, this means there is low similarity between the previous and current location. On the other hand small difference implies high similarity which increases the weight as in (5.13).

5.3 Rotation of Action Direction

In order to use velocity information in action recognition it is essential to convert all the videos into same type of motion direction. To do this, the velocity components of GM-PHD state vectors are transformed into the reference direction.

The rotation process is performed in the way that global motion vector of the action is extracted and the velocity and covariance components of PHD filter are transformed to the reference direction. The rotation process is performed in the way that global motion vector of the action is extracted and the velocity and covariance components of PHD filter are transformed to the reference direction. The reference direction is chosen to be the right direction and transformation is performed by finding the global motion vector and making 2D transformation of the velocity vectors.

Average velocity of the over-all surface for the j^{th} video is found by taking the

velocity components of the average state estimate vector in (5.14).

$$m_{avg}^{j} = \frac{1}{M_j} \sum_{k=1}^{M_j} \frac{1}{N_j(k)} \sum_{i=1}^{N_j(k)} m_i^j(k) \tag{5.14}$$

where $m_i^j(k)$ stands for the mean vector for the i^{th} Gaussian in the estimated PHD and , N_j and M_j corresponds to the number of Gaussians in the PHD at time k and number of frames for the j^{th} video, respectively. This vector in (5.15) can also be represented in the polar coordinate system as in (5.16). Direction of motion is found by simply taking the angle of the mean velocity vector Θ_v^j.

$$m_{v,avg}^{j}(k) \triangleq [m_{v_x,avg}^{j}(k)\ m_{v_y,avg}^{j}(k)]' \tag{5.15}$$

$$= \left(\underbrace{\sqrt{m_{v_x,avg}^{j}(k) + m_{v_y,avg}^{j}(k)}}_{R_v^j}, \underbrace{\tan^{-1} m_{v_y,avg}^{j}(k)^2 / m_{v_x,avg}^{j}(k)^2}_{\Theta_v^j} \right) \tag{5.16}$$

Then all the velocity components of each videos are transformed in order to bring all the velocity directions towards the $+x$ direction in average.

$$\widetilde{m}_v^j(k) = \begin{bmatrix} \cos(\Theta_v^j) & \sin(\Theta_v^j) \\ \sin(-\Theta_v^j) & \cos(\Theta_v^j) \end{bmatrix} m_v^j(k) \tag{5.17}$$

where $j = 1, \ldots, N_{total}$ and $m_v^j(k) \triangleq [m_{v_x}^j(k)\ m_{v_y}^j(k)]'$.

5.4 High Level Recognition by Global Motion Analysis

When we analyze the characteristics of the video, we observe that we can define a global motion on differentiation on running, jogging and walking actions from the boxing, handclapping and handwaving actions. The analysis of this global motion makes a high level recognition possible. In High Level Recognition, we categorize the actions into two groups:

- **Group1:** Running, Jogging and Walking actions.

- **Group2:** Boxing, Handwaving and Handclapping actions.

In High Level Recognition, x velocity of the motion, $m^j_{\widetilde{v_x}}(k)$, is utilized for classification purposes. The actions which has average x velocity greater than a certain threshold is chosen to be in Group-1 which indicates a global motion towards x direction. If x velocity is less than a threshold, the action is chosen to be in Group-2 as given in (5.18).

$$j^{th} \text{ Video} \in \begin{cases} Group-1, & \text{if } \bar{m}^j_{v_x} \geq Th \text{ px/frame.}, \\ Group-2, & \text{otherwise.} \end{cases} \tag{5.18}$$

where Th indicates the threshold value.

5.4.1 Determination of the Threshold Value

In order to make a classification with x velocity component, $m^j_{\widetilde{v_x}}(k)$, in an efficient way, first we have to eliminate outliers. The outliers are the points belonging to the background and have lower velocity values that decrease the average velocity and make the discrimination erroneous. It is certain that zooming operations results in a considerable velocity for background observations, yet this effect is ignored for the sake of simplicity. In the first step, we eliminate outliers using position information of the Gaussian Mixture intensity. In elimination, the GM-PHD modes which are not in an interval determined by standard deviation is discarded. For this purpose, mean and standard deviation of position of GM-PHD modes are calculated at each frame as given (5.19) and (5.20). Then the modes which satisfies the conditions given in (5.21) and (5.22) are selected for High Level Recognition purposes as given in (5.23).

$$\bar{m}^j_x(k) = \frac{1}{N^j(k)} \sum_{i=1}^{N^j(k)} m^j_{i,x}(k) \tag{5.19}$$

$$\sigma_x^j(k) = \sqrt{\frac{1}{N^j(k)} \sum_{i=1}^{N^j(k)} \left(m_{i,x}^j(k) - \bar{m}_x^j(k)\right)^2} \tag{5.20}$$

$$cond_1(i) = \left|m_{i,x}^j(k) - \bar{m}_x^j(k)\right| < 1.2 * \sigma_x^j(k) \tag{5.21}$$

Similarly, mean and standard deviation for the y components are found and another condition is obtained as in (5.22).

$$cond_2(i) = \left|m_{i,y}^j(k) - \bar{m}_y^j(k)\right| < 2 * \sigma_y^j(k) \tag{5.22}$$

$$i^{th} \text{ Mode} \begin{cases} \in S^j(k), & \text{if } cond_1 \text{ and } cond_2 \text{ holds,} \\ discard, & \text{otherwise.} \end{cases} \tag{5.23}$$

After outlier elimination, HLR threshold value is calculated by using mean of transformed x velocity information, $m_{i,\widetilde{v_x}}^j(k)$, over L frames as given in (5.24).

$$\bar{m}_{\widetilde{v_x}}^j = \frac{1}{L} \sum_{k=1}^{L} \frac{1}{N_S^j(k)} \sum_{i \in S^j(k)} m_{i,\widetilde{v_x}}^j(k) \tag{5.24}$$

where N_S represents the number of elements in the set $S^j(k)$ and L is selected as a constant value which is 15.

Then the threshold for HLR operation is chosen by analyzing the mean, $\bar{m}_{\widetilde{v_x}}^j$, of all videos. The histogram of mean of transformed x velocity components over 15 frames,$\bar{m}_{\widetilde{v_x}}^j$, of train videos over 15 frames is given in Figure 5.1.

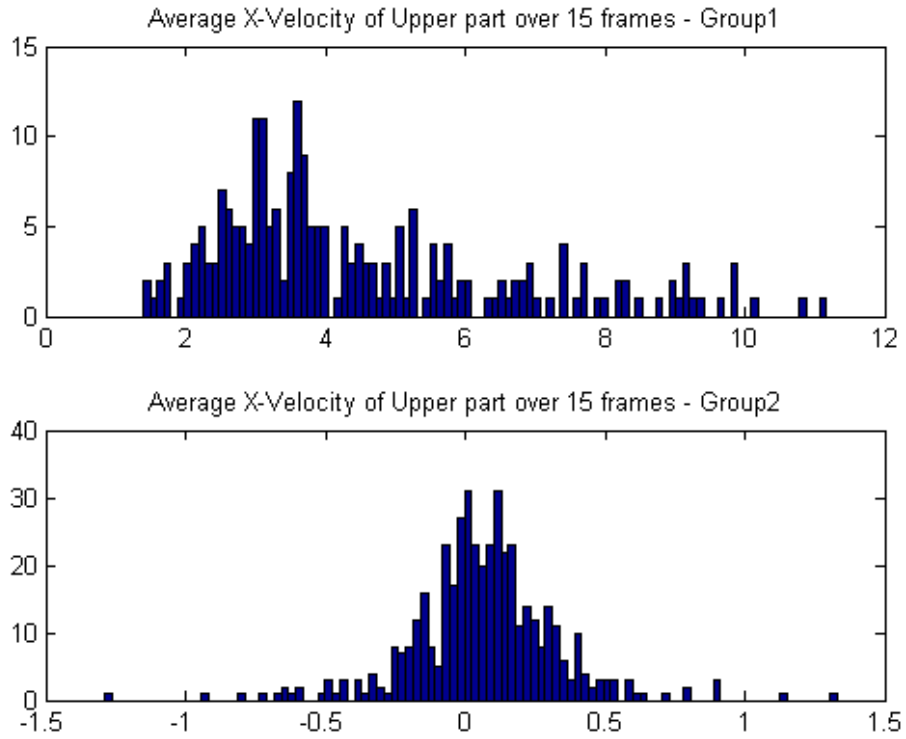


Figure 5.1: Histogram of transformed X-Velocity state of GM-PHD

The histogram of, $\bar{m}^j_{\widetilde{v_x}}$, reveals us choosing the threshold value as 1.4 makes HLR possible for GR-1 and GR-2 actions.

5.4.2 Selection of the Group

After determining the threshold value as 1.4 px/frame, the selection of group is simply performed as in (5.25).

$$j^{th} \text{ Video} \in \begin{cases} Group-1, & \text{if } \bar{m}^j_{v_x} \geq 1.4 \text{ px/frame.}, \\ Group-2, & \text{otherwise.} \end{cases} \tag{5.25}$$

5.5 Determination of Discriminative Modes of GM-PHD

After performing HLR, the way of finding discriminative modes providing valuable information is a critical step for recognition. The discriminative mode selection is simply choosing the Gaussians of GMPHD which give information about the foreground object by eliminating the background information. The outliers are the points that do not belong to the person in action. In the outlier elimination process, we assume that most of the observations in the frame are localized on the person taking the action.

The outliers are randomly distributed over the video frame and it may be quiet difficult to estimate these points. One of the basic assumptions is that state estimates belonging to the outliers are far away from the states belonging to the body and their velocity is much slower than that of the body state estimates.

The selection of discriminative modes are performed in each frame and is performed for both Group-1 and Group-2 action sets in different manners since the action characteristic is different for each group.

5.5.1 Group-1

In Group-1 action set, we find the mean and standard deviation of the x and y positions of the GM-PHD state estimates as given in (5.27). Then we eliminate the observations whose positions away from the mean with a constant multiple of standard deviation given in (5.28) and (5.29).

$$\bar{m}_x^j(k) = \frac{1}{N^j(k)} \sum_{i=1}^{N^j(k)} m_{i,x}^j(k) \tag{5.26}$$

$$\sigma_x^j(k) = \sqrt{\frac{1}{N^j(k)} \sum_{i=1}^{N^j(k)} \left(m_{i,x}^j(k) - \bar{m}_x^j(k)\right)^2} \tag{5.27}$$

$$cond_1(i) = \left|m_{i,x}^j(k) - \bar{m}_x^j(k)\right| < 2 * \sigma_x^j(k) \tag{5.28}$$

Similarly, mean and standard deviation for the y components are found and another condition is obtained.

$$cond_2(i) = \left| m^j_{i,y}(k) - \bar{m}^j_y(k) \right| < 2 * \sigma^j_y(k) \tag{5.29}$$

The outlier elimination process is very similar to HLR but the threshold values. In outlier elimination using position information, the thresholds are chosen to to be wider compared to HLR since in HLR a narrower window is more appropriate in order to take the modes close to body center is sufficient. But in GR-1 we widen the threshold window in order not to miss the information belonging to body in action.

In addition to elimination of outliers with respect to position information, the GM-PHD modes which have lower velocity magnitudes then 0.5 are also eliminated. This elimination is performed assuming the fact that there is always a global displacement of the body in Group-1 videos which result in a considerable velocity of the modes in average. Besides since the legs are opened wide, the standard deviation is not able to eliminate all outliers in the leg part. To eliminate the GM-PHD modes that have slow motion component, we discard the features having a speed value less than a constant as given in (5.30). In this process, one has to be aware the fact that one of the legs stay almost stationary and we except to lose information regarding that part of the body.

$$cond_3(i) = \sqrt{\widetilde{m}^j_{i,v_x}(k)^2 + \widetilde{m}^j_{i,v_y}(k)^2} > 0.5\text{px/frame} \tag{5.30}$$

$$i^{th} \text{ Mode} \begin{cases} \in S^j_{\text{GR-1}}(k), & \text{if } cond_1, \, cond_2 \text{ and } cond_3 \text{ holds,} \\ discarded, & \text{otherwise.} \end{cases} \tag{5.31}$$

where $S^j_{\text{GR-1}}(k)$ represents the selected modes of Gaussian Mixture PHD. Modes that do not satisfy all the conditions $cond_1$, $cond_2$ and $cond_3$ at the same time are assumed to be belonging to the background and are eliminated.

5.5.2 Group-2

In Group-2 action set, the lower part stands almost the same although the upper part of the body changes as the hands and arms make clapping, waving and boxing actions. Since the position information of the lower part is more stable, we perform elimination of outliers using the information obtained from the lower part of the body as basis. In the first step, we find lower part, as explained in 5.6, of the person and eliminate the outliers using the mean and standard deviation information calculated as (5.33). position of the this information as given in (5.35).

Then we eliminate the features whose positions away from the mean with a constant multiple of standard deviation along only x direction as given in (5.34).

$$\bar{m}_x^j(k) = \frac{1}{N^j(k)} \sum_{i=1}^{N^j(k)} m_{i,x}^j(k) \tag{5.32}$$

$$\sigma_x^j(k) = \sqrt{\frac{1}{N^j(k)} \sum_{i=1}^{N^j(k)} \left(m_{i,x}^j(k) - \bar{m}_x^j(k)\right)^2} \tag{5.33}$$

$$cond_1(i) = \left|m_{i,x}^j(k) - \bar{m}_x^j(k)\right| < 1.5 * \sigma_x^j(k) \tag{5.34}$$

$$i^{th} \text{ Mode} \begin{cases} \in S_{\text{GR-2}}^j(k), & \text{if } cond_1 \text{ holds,} \\ discarded, & \text{otherwise.} \end{cases} \tag{5.35}$$

The outlier elimination approach is very similar to GR1 but the threshold values and body part. In outlier elimination using position information, the thresholds are chosen to be narrower compared to GR1 since we make elimination considering only lower part of the body which do not have motion in general since the body stands in GR-2 actions.

5.6 Clustering

In KTH database actions, the lower and upper part motion characteristics of the actions are different for both GR-1 and GR-2 actions. So, we choose to analyze them separately. After eliminating the outliers and determining the discriminative modes of GMPHD as stated in 5.5, we cluster the GMPHD modes using position information. The clustering is performed for each group in different manner. For GR1 actions, the body is clustered into 2 as Upper Part(UP) and Lower Part(LP) of the body, where as for GR-2 actions, the clustering is made into 3 as Upper Left(UL), Upper Right(UR) and LP of the body. This clustering approach is constructed based on body motion of each cluster, since the legs are wide open in GR1 actions, we prefer to analyze them seperately. In GR-2 actions, although there is slight motion is leg part, discriminative information is in upper part which has a symmetric motion in hand-clapping and hand-waving actions which sum up to 3 clusters for GR-2.

Firstly, we mentioen about determination of upper and lower parts, then UL and UR part extraction is explained.

5.6.1 Determination of Upper and Lower parts for GR-1 and GR-2

After extracting discriminative modes of GMPHD and eliminate outliers, the determination of Upper and Lower parts is done by simply dividing the body horizontally into two with 60 percent ratio as given in (5.37). First the height of the person is calculated using the upper and lower positions along y axis as given in (5.36).

$$height = \max_i m^j_{i,y}(k) - \min_i m^j_{i,y}(k), i \in S^j_{\text{GR-1}}(k) \tag{5.36}$$

We assume that upper and lower parts correspond to head/torso and legs, respectively. To find the upper and lower part we perform a simple clustering with 60% ratio. This ratio is selected after analyzing the body motions. We assumed that leg part of the body consists of 40 percent of the body and performed the clustering as stated in (5.37).

$$i^{th} \text{ Mode} \begin{cases} \in S^j_{\text{GR-1,UP}}(k), & \text{if } m^j_{i,y}(k) - \min_i m^j_{i,y}(k) > 0.6 * height, \\ \in S^j_{\text{GR-1,LP}}(k), & \text{otherwise.} \end{cases} \tag{5.37}$$

Similar operations are performed for Group-2 and the body is clustered as Lower and Upper part.

5.6.2 Determination of Upper-Left and Lower-Left parts for GR-2

For Group-2 actions, it is observed that there exists symmetricity with respect to the central vertical axis for most of the people in the videos, while this is not the case for boxing. So we divide the upper part into 2 clusters using the x center of the lower part of the body which is calculated as given in (5.38). Since the modes of legs fits constant velocity model and more robust since there is few motions, we prefer to take lower part as reference. and for which the dividing axis is obtained as the x position mean of the lower part as given in (5.38) and (5.39).

$$\bar{m}_{x_L} = \frac{1}{N^j_{GR-2,LP}(k)} \sum_{i \in S^j_{\text{GR-2,LP}}(k)} m^j_{i,x}(k) \tag{5.38}$$

where $N^j_{GR-2,LP}(k)$ corresponds to the number of GM-PHD modes in lower part of Group-2.

The Upper Left and Upper Right discrimination is performed as stated in (5.39)

$$i^{th} \text{ Mode} \begin{cases} \in S^j_{\text{GR-2,UL}}(k), & \text{if } m^j_{i,x}(k) \in S^j_{\text{GR-2,UP}}(k) < mean_{x_L}, \\ \in S^j_{\text{GR-2,UR}}(k), & \text{otherwise.} \end{cases} \tag{5.39}$$

5.7 Construction of HMM Structure

In this study, we extract Harris corner observations in each frame and perform group tracking for these observations by the GM-PHD filter. GM-PHD make

possible to perform group tracking for these features without any association process and eliminating the features that does not fit to the defined motion model. Besides it is capable of tracking the occluded observations. The action recognition is performed by utilizing the Hidden Markov Model technique and a separate HMM structure is constructed for every action. These structures include independent HMMs for every feature parameter. HMM features are extracted from the GM-PHD filter output and are given in Table 5.2. Note that this feature set is defined for every mode of GM-PHD in each frame of each video.

Table 5.2: GM-PHD outputs used as a basis for HMM

Parameter	Explanation
m_x	X-Position
m_y	Y-Position
m_{v_x}	X-Velocity
m_{v_y}	Y-Velocity
w	Weight
P_p	Position Covariance
P_v	Velocity Covariance

Number of observations in each frame may be up to 100 and it is impossible to propose a HMM model for each estimate. For reduction of the information, mean values of selected information captured by different clusters in a single frame are found. Besides, GM-PHD difference and OSPA difference are found for the frames belonging to different time instants. The complete feature vector set used in the HMM structure for each cluster is provided in 5.3.

Note that there are 10 feature parameters for each cluster. If we consider Group1 HMM structure, there are 2 clusters, upper and lower parts, in each frame. This means there are 20 independent HMM's for every action (running, jogging, walking).

For Group2 actions, we divide the body into 3 clusters composed of legs(lower part), upper left and upper right portions of the body. This means there are 30 independent HMM's for every Group2 action (boxing, handwaving, handclap-

Table 5.3: Feature parameter vector for HMM

Parameter	Explanation
$\bar{m}_x$	mean of X-Positions
$\bar{m}_y$	mean of Y-Positions
$\bar{m}_{v_x}$	mean of Speed
$\bar{m}_{v_y}$	mean of Velocity Directions
σ_{m_x}	std of X-Position
σ_{m_y}	std of Y-Position
σ_{m_r}	std of Speed
σ_{m_Theta}	std of Velocity Direction
GM-PHD distance	Distance of GMPHD between consecutive frames
OSPA Distance	OSPA metric between consecutive frames

ping).

In the following subsections, we describe the way of constructing each action recognition feature parameter in a detailed manner for obtaining HMM Feature Set.

5.7.1 HMM Structure

In this section the properties of HMM in use will be defined. A Hidden Markov Model (HMM) is a Bayesian network in which the modeled system is to be a Markov process. In the action recognition problem in which the input is a video sequence, the correlation between the video frames contains critical information for discrimination of an action. The sequence of movements of the body is action specific and plays a crucial role in action recognition by visual sources. In this thesis, we use Hidden Markov Model in order to utilize the information between the sequences of movements. HMM is a Model which assumes the input given to the model is a Markov process which fits to video action recognition problem well. Considering the solution for a human, when the human knows the previous posture of the body, the movement and the following posture can be estimated. The same idea serves a basis for utilizing HMM in action recognition in solving the problem. In HMM, the relation in-between the sequences of information

is obtained and hold in by the HMM parameters in the training phase. Then, in action recognition phase, the most probable HMM is found and the action owning this HMM is chosen as the recognition result.

As stated before, we use independent HMM structures for each action and each HMM structure contains one HMM for each feature parameter. After the group of the action is determined in the High Level Recognition phase, then different HMM structures are constructed for Group1 and Group2 actions which are provided in Table 5.4 and 5.5.

Table 5.4: Group-1 : Running, Walking and Jogging action properties

Parameter	Value
No. of Clusters	2
Feature For each Cluster	10
General Parameters	2
Total No. of Paramaters	22

Table 5.5: Group-2 : Boxing, Hand-Clapping and Hand-Waving action properties

Parameter	Value
No. of Clusters	3
Feature For each Cluster	10
General Parameters	2
Total No. of Paramaters	32

Notice that in the HMM structure of every action type, there are independent HMM's for each HMM parameter which sums up to 22 and 32 number of HMM's for Group-1 and Group-2, respectively.

Talking about the HMM properties, the same HMM frame is used for each feature of each action which is given in Figure 5.2. The properties of each HMM frame is given in 5.6. For a detailed information about the properties, HMM mechanism is explained in 2.3. To mention shortly, Hidden State number is chosen to be 4 which yields a 1x4 prior probability matrix and a 4x4 state transition matrix. During the experimentation phase of the thesis, we realize

that the hidden number of states, 4, is well sufficient for recognition purposes for 3 videos. Also note that, we performed some experiments for 3 different state-transition matrix mechanisms which is mentioned in 6.3.1 in detail.

Considering the representation of outputs, the code-book generation is performed by quantized values of the HMM parameters. So for the code-book generation i.e., quantization of the HMM parameters is performed and observation symbols are generated. The calculation of each HMM feature parameter and data preparation processes are mentioned in Section 5.7.2. Briefly, in HMM data preparation phase, every feature is quantized and mapped to 1-11 interval which directly corresponds to the HMM output states as mentioned in 5.7.2. As a result, we construct 11 observation symbols whose values can take the integers between 1 and 11. Since there are 11 observation symbols and 4 hidden models, the dimension of the observation matrix is of 4x11.

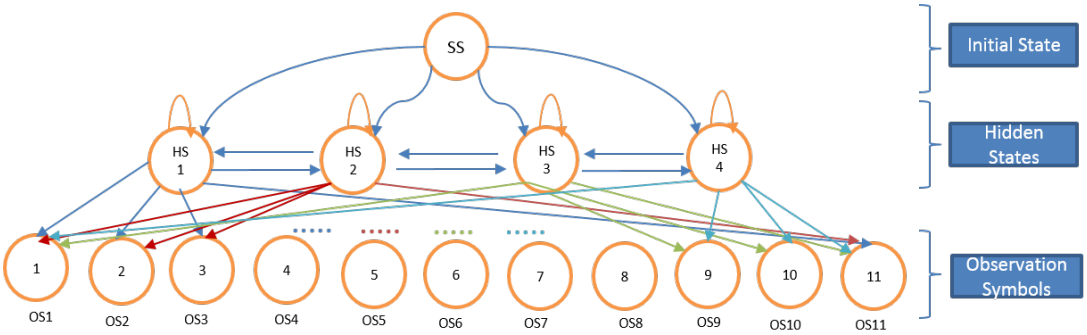


Figure 5.2: HMM Frame for each feature of each action

The training phase is performed for the 40% of the videos in the KTH database, and testing phase is performed with the rest. In the testing phase, after performing High Level recognition, every sequence belonging to each action is tested by the trained HMM's of each action. Then the best fitting class of the related feature is found by taking the class giving out the maximum Posterior Probability. This recognition process is summarized in Figure 5.3.

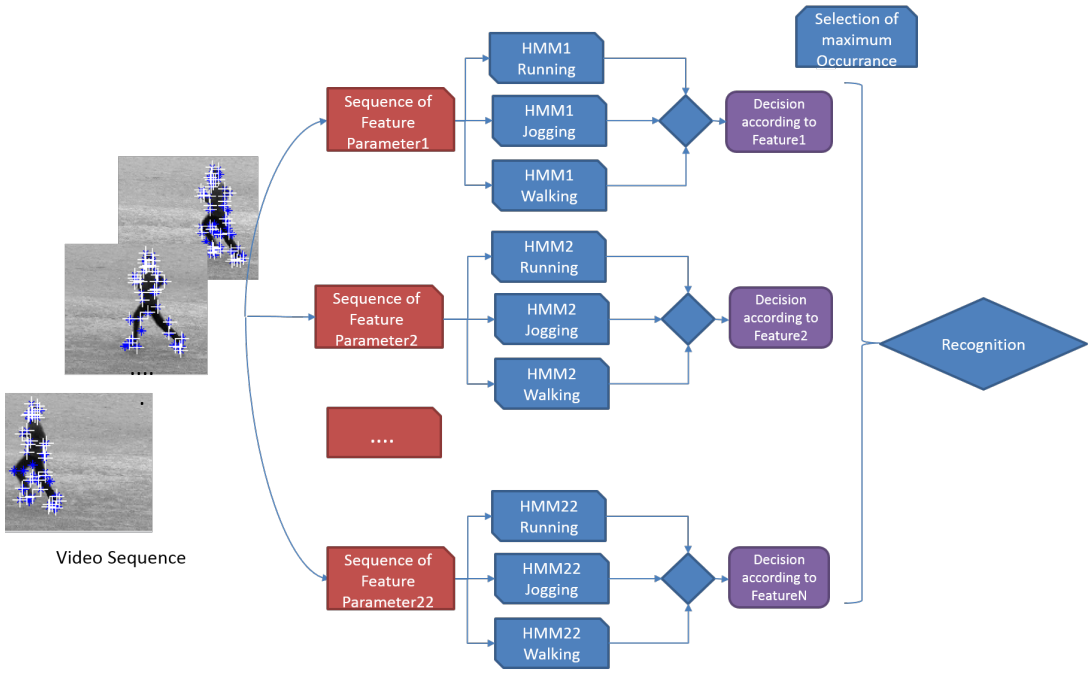


Figure 5.3: Recognition by HMM

In order to prevent getting stuck to a local maxima and to find the global maxima, we make 20 iterations for the selection of starting point. Best selection is the point providing the highest likelihood as a result. Then, the resultant HMM variables are taken for this selection. The convergence criterion for the HMM is selected to be 0.0001 and maximum iteration number of Baum-Welch algorithm is selected to be 20.

In recognition phase, every feature sequence of each video is evaluated in 3 independent HMMs of different actions (Running,Jogging,Walking for Group-1, Boxing, HandClapping, HandWaving for Group-2). Then the posterior probabilities are calculated and the HMM giving the maximum posterior probability is selected to be the action belonging to that specific feature. This process is performed for 22 and 32 features for Group-1 and Group-2, respectively. Then we obtain a 22x1 and 32x1 matrices which guide us for recognition. Finally, the recognition of action is performed by selecting the action with maximum occurrence in Feature Decision Vector as given in Tables 6.8 and 6.10.

Table 5.6: HMM properties

Parameter	Value
Hidden State	4
Observation Symbol	11
Prior Probability Matrix	1x4
State Transition Probability Matrix (3 different type)	4x4
Emission Probability Matrix	4x11
Iteration Starting point	20
Convergence criterion	1e-4
Max iteration number	20
Training algorithm	Baum-Welch
Testing (Classification) Criterion	Maximum posterior probability, Max(class)

$$\text{Feature Decision Vector(FDV)} = [\text{ActionDec1}, \text{ActionDec2}, \ldots \text{ActionDec22}] \tag{5.40}$$

$$\text{No. of Occurrences of FDV} = [\text{No. of Act. Type1}, \text{No. of Act. Type2}, \text{No. of Act. Type3}] \tag{5.41}$$

$$\text{Class} = \max [\text{No. of Act. Type1}, \text{No. of Act. Type2}, \text{No. of Act. Type3}] \tag{5.42}$$

5.7.2 Construction of HMM Feature Set

In this section, the way of constructing HMM feature set is explained. As stated before, Harris corner detector extracts observations for both background and foreground scene. Based on the assumption that most of the observations are localized on the human body, we eliminate the features belonging to the background by some kind of gating using the standard deviation of the overall information. For a consistent action recognition, it is also necessary to eliminate

the effects which cause a scaling difference in-between different videos and to map all the HMM features to a predefined reference coordinate system. In this section all the operations which are performed to bring the features to the same reference is defined for every HMM feature. Detailed information on feature extraction phase for HMM is provided and explanations regarding how we obtain the feature parameter values used in HMM for action recognition purposes are given.

5.7.2.1 Mean and Standard Deviation of Position

Human recognizes the action by looking at the person body and limb motion without considering the background. In parallel to this statement, in action recognition algorithms, not the position of the feature in the frame but the relative position of the features with respect to the body center is critical. Another important observation effecting the position of the features is that the height of the person is different in each frame. This is because of the fact that every person has different height and the distance between the person and the camera may be different. We propose a solution for the scaling operation to bring all types of video sequences to the same reference.

In the first step, the uppermost, lowermost, rightmost and leftmost x and y position limits of are found separately. Then all the positions are mapped to 1-11 interval as given in (5.44).

$$Y^j_{Range}(k) = \max_i m^j_{i,y}(k) - \min_i m^j_{i,y}(k),\ i \in S^j_{\text{GR-1,UP}}(k) \tag{5.43}$$

$$X^j_{Range}(k) = \max_i m^j_{i,x}(k) - \min_i m^j_{i,x}(k),\ i \in S^j_{\text{GR-1,UP}}(k) \tag{5.44}$$

$$m^j_{q,i,y}(k) = (m^j_{i,y}(k) - \min_i m^j_{i,y}(k))/Y^j_{Range}(k) * 10 + 1, i \in S^j_{\text{GR-1,UP}}(k) \tag{5.45}$$

$$m^j_{q,i,x}(k) = (m^j_{i,x}(k) - \min_i m^j_{i,x}(k))/X^j_{Range}(k) * 10 + 1, i \in S^j_{\text{GR-1,UP}}(k) \tag{5.46}$$

The equation is given for the Upper part of GR-1 actions. Same operations are performed for all 3 clusters of GR-2 and lower part of GR-1.

After mapping the values to the interval [1-11], mean and std of the x-position and y-positions are found for each cluster. Finally, quantization is performed by the rounding operation and the HMM feature parameter values related to the positions are obtained as in (5.48).

$$mean^{j}_{m_{q,x}}(k) = \text{round} \frac{1}{N^{j}_{GR-1,UP}(k)} \sum_{i \in S^{j}_{\text{GR-1,UP}}(k)} m^{j}_{q,i,x}(k) \tag{5.47}$$

$$mean^{j}_{m_{q,y}}(k) = \text{round} \frac{1}{N^{j}_{GR-1,UP}(k)} \sum_{i \in S^{j}_{\text{GR-1,UP}}(k)} m^{j}_{q,i,y}(k) \tag{5.48}$$

$$\sigma^{j}_{m_{q,x}}(k) = \text{round} \sqrt{\frac{1}{N^{j}(k)} \sum_{i=1}^{N^{j}(k)} \left(m^{j}_{q,i,x}(k) - \bar{m}^{j}_{q,x}(k) \right)^2} \tag{5.49}$$

$$\sigma^{j}_{m_{q,y}}(k) = \text{round} \sqrt{\frac{1}{N^{j}(k)} \sum_{i=1}^{N^{j}(k)} \left(m^{j}_{q,i,y}(k) - \bar{m}^{j}_{q,y}(k) \right)^2} \tag{5.50}$$

5.7.2.2 Mean and Standard Deviation of Speed

The distance between the person and the camera varies both within the video frames and within different types of videos. As a result, the height of the person observed for both within a video sequence and for different videos varies. In order to overcome this problem, calculated height of the person in action is mapped to a reference height. The diversity of the distance of the person to the camera also affects the velocity components. In order to work in the same scale the velocity is also transformed to a reference.

Proposed GMPHD state definition includes velocities along x and y directions. By taking the velocity components of the state estimates into account, speed information is obtained as in equation (5.16).

Since speed is extracted using 2D image domain, speed of the person changes as

the position of the person to the camera changes. So, if we want to use speed information in action recognition, it is necessary to use the same speed scale in all videos.

In order to obtain the reference, we perform a post processing operation and obtain the height range of the person from all videos given in Figure 5.4 and the range is selected as $Ref_{Height} = 110$.

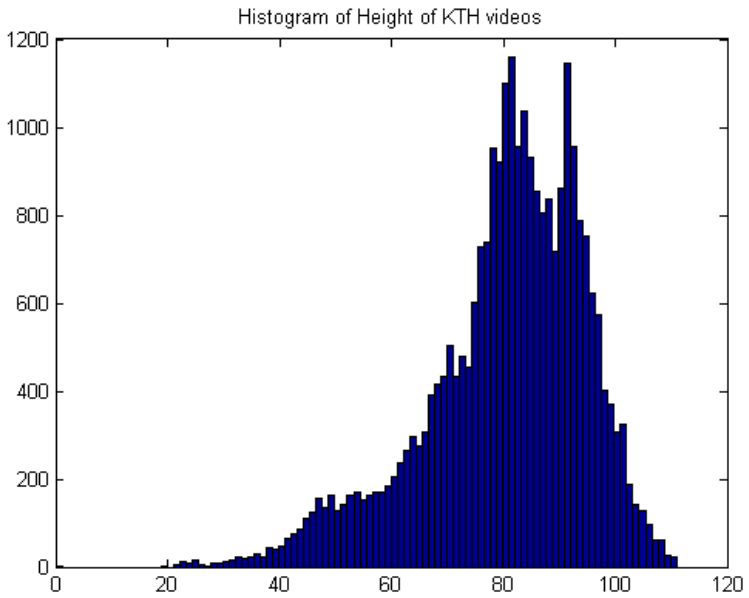


Figure 5.4: Histogram of Height of person in action in KTH database.

The feature set related to speed information is obtained by:

i Person height in the frame is found by taking difference upper and lower Y position as given in (5.36).

ii Speed is transformed to a domain independent from person height and position with respect to the camera.

$$R^j_{v_i,m}(k) = R^j_{v_i}(k) * Ref_{Height}/Height \tag{5.51}$$

iii Speed value is delimited to 0-15 ($Speed_{Limit} = 15$) interval after analyzing the speed by obtaining speed histogram given in Figure 5.5.

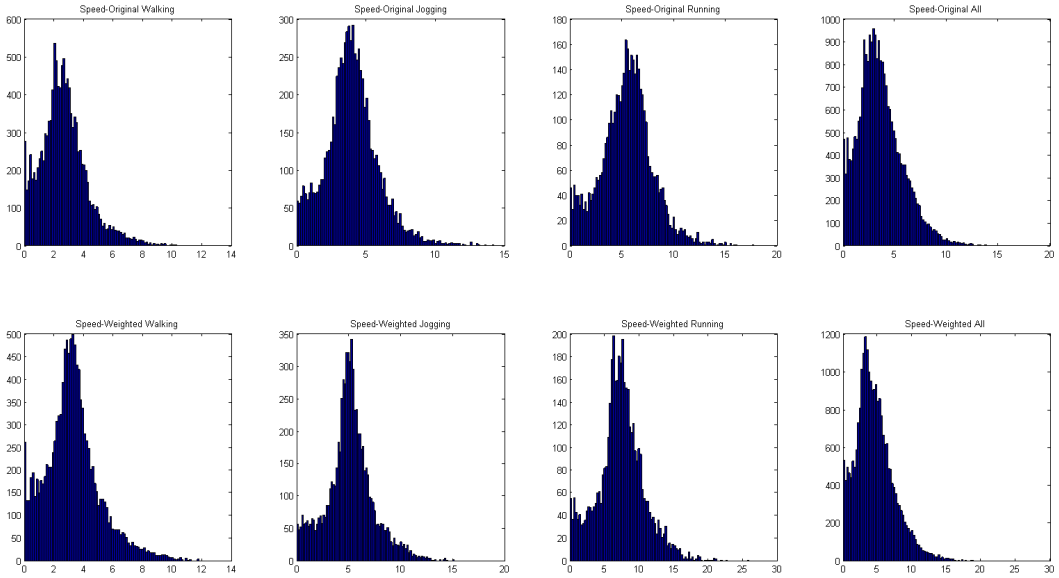


Figure 5.5: Histogram of Speed of person in action in KTH database.

Then the delimited speed is mapped to 1-11 interval as given in (5.52)

$$R^j_{v_i,q}(k) = R^j_{v_i,m}(k)/Speed_{Limit} * 10 + 1 \tag{5.52}$$

iv Mean and std. of speed vectors is found for each cluster as given in (5.53) and (5.54)

$$\bar{R}^j_{v_i,q}(k) = \text{round}\frac{1}{N^j_{GR-1,UP}(k)} \sum_{i \in S^j_{\text{GR-1,UP}}(k)} R^j_{v_i,q}(k) \tag{5.53}$$

$$\sigma_{R^j_{v_i,q}(k)} = \text{round}\sqrt{\frac{1}{N^j(k)} \sum_{i=1}^{N^j(k)} \left(R^j_{v_i,q}(k) - \bar{R}^j_{v_i,q}(k)\right)^2} \tag{5.54}$$

Mean and standard deviation of speed information belonging to GR-1 lower part and all parts of GR-2 is calculated in the same manner.

5.7.2.3 Mean and Standard Deviation of Angle

When we consider the direction of body motion, the first prominence is that it changes between videos. In other words, the body moves along different directions. In order to use velocity angle for action discrimination purposes, we need to map the angle to a reference direction.

i Velocity Angle is transformed towards the right direction

In order to construct a reference, global motion vector of the motion is extracted and the velocity is transformed to the reference direction as explained in 5.3.

In analyzing Velocity Angle, we investigated the correlation of speed with angle. We performed this analysis since in HMM recognition they are taken as independent. If there are any correlation we might have miss the information. So, the velocity speed vs. angle map is constructed for GR-1 actions to visualize the relationship which is given in Figure 5.6.

For a more detailed analysis, histogram is obtained by dividing the body into 3 portions which correspond to head,torso and foot respectively. The figure indicates that there is no direct correlation between the components which make us decide to use speed and angle information independently.

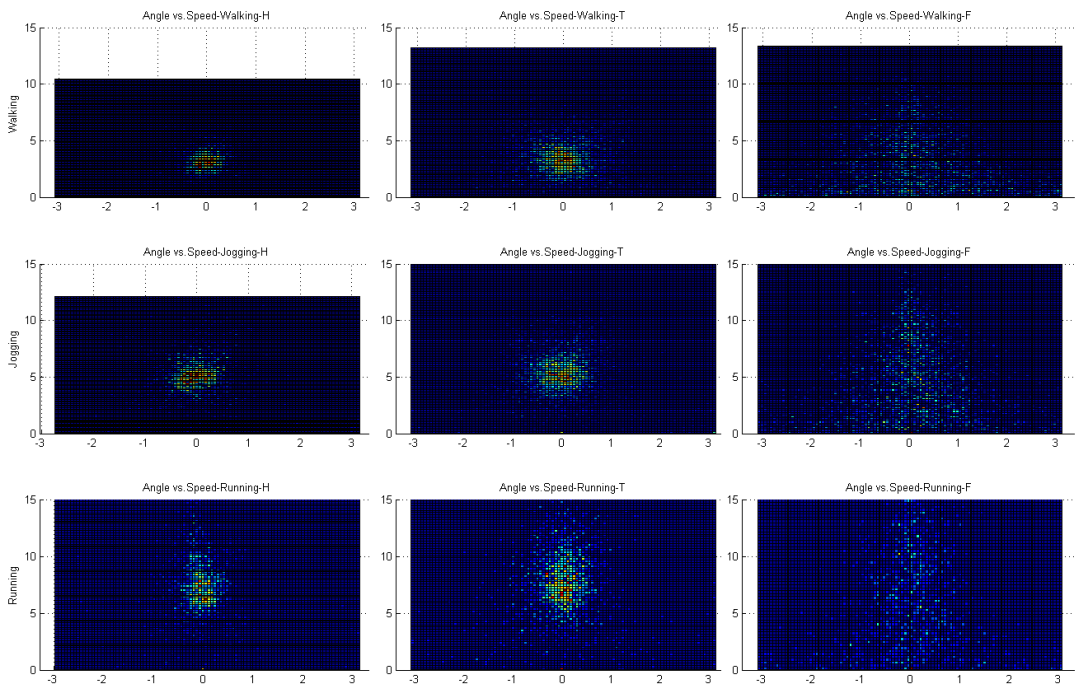


Figure 5.6: Histogram Speed vs Velocity Angle of GR-1 actions

ii Velocity angle limits are taken as to cover 2 *pi range after analyzing the velocity direction by obtaining angle histogram given in Figure 5.7.The analysis performed for GR-1 actions since it is obvious that hand-waving operation covers 2π range. So following values for the related parameters are taken in the algorithms:

$$Angle_{Limit} = \pi \tag{5.55}$$

$$Min_{angle} = \pi \tag{5.56}$$

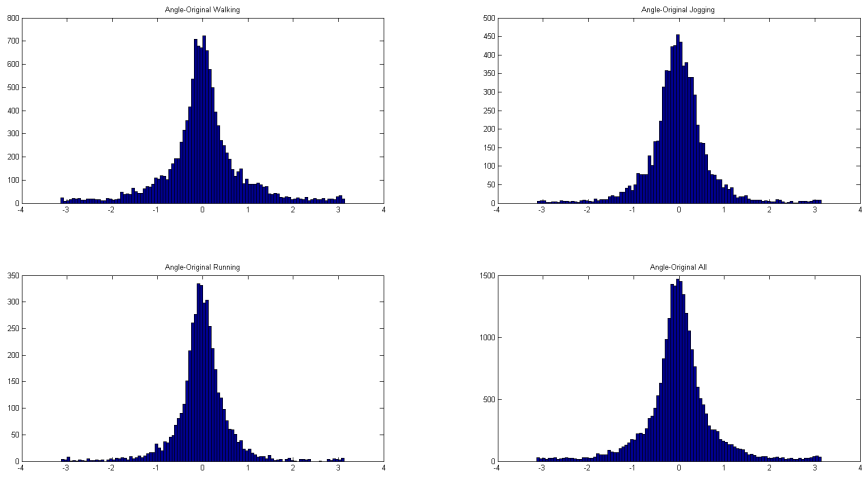


Figure 5.7: Histogram Velocity Angle of GR-1 actions

Then the delimited speed is mapped to 1-11 interval as given in (5.57)

$$\Theta^{j}_{v_i,q}(k) = \Theta^{j}_{v_i,m}(k) - Min_{angle}/Angle_{Limit} * 10 + 1 \tag{5.57}$$

iii Mean and std. of velocity angle vectors are found for each cluster as given in (5.58) and (5.59)

$$\bar{\Theta}^{j}_{v_i,q}(k) = \text{round}\frac{1}{N^{j}_{GR-1,UP}(k)} \sum_{i \in S^{j}_{\text{GR-1,UP}}(k)} \Theta^{j}_{v_i,q}(k) \tag{5.58}$$

$$\sigma_{\Theta^{j}_{v_i,q}(k)} = \text{round}\sqrt{\frac{1}{N^{j}(k)} \sum_{i=1}^{N^{j}(k)} \left(\Theta^{j}_{v_i,q}(k) - \bar{\Theta}^{j}_{v_i,q}(k)\right)^2} \tag{5.59}$$

Mean and standard deviation of velocity angle information belonging to GR-1 lower part and all parts of GR-2 is calculated in the same manner.

5.7.2.4 GM-PHD Distance

GM-PHD distance is taken as the difference between the Gaussian mixtures belonging to different time instants. This distance corresponds to the change in

the Gaussian mixture functions, including all positional, velocity, weight and covariance information in time. In this thesis, the distance is taken for consecutive frames. Definition and calculation of this distance is provided in this section.

Remembering the GM-PHD is a Gaussian Mixture function as given in (5.60), the function to find the distance between PHD's belonging to different time instants,i.e, k_1 and k_2, is proposed as in (5.61).

$$p(x,k) = \sum_{i=1}^{N(k)} w_i(k)\mathcal{N}(x; m_i(k), P_i(k)) \tag{5.60}$$

$$D(k_1,k_2) = \int (p(x,k_1) - p(x,k_2))^2 dx \tag{5.61}$$

If the subtraction in (5.61) is focused, another Gaussian mixture function is obtained.

$$p(x,k_1) - p(x,k_2) = \sum_{n=1}^{N(k_1)+N(k_2)} w_n(k_1,k_2)\mathcal{N}(x; m_n(k_1,k_2), P_n(k_1,k_2)) \tag{5.62}$$

where

$$w_n(k_1,k_2) = \{w_1(k_1), \ldots, w_{N(k_1)}(k_1), -w_1(k_2), \ldots, -w_{N(k_2)}(k_2)\}_{n=1}^{N(k_1)+N(k_2)} \tag{5.63}$$

$$m_n(k_1,k_2) = \{m_1(k_1), \ldots, m_{N(k_1)}(k_1), m_1(k_2), \ldots, m_{N(k_2)}(k_2)\}_{n=1}^{N(k_1)+N(k_2)} \tag{5.64}$$

$$P_n(k_1,k_2) = \{P_1(k_1), \ldots, P_{N(k_1)}(k_1), P_1(k_2), \ldots, P_{N(k_2)}(k_2)\}_{n=1}^{N(k_1)+N(k_2)} \tag{5.65}$$

$$\tag{5.66}$$

Taking the square of the Gaussian Mixture function in (5.61), (5.67) is found.

$$(p(x,k_1) - p(x,k_2))^2 = \sum_{n=1}^{N(k_1)+N(k_2)} \sum_{p=1}^{N(k_1)+N(k_2)} w_n w_p \mathcal{N}(x; m_n, P_n)\mathcal{N}(x; m_p, P_p) \tag{5.67}$$

(5.67) includes multiplication of two Gaussian functions which results in another Gaussian multiplied by a scalar, (5.68).

$$\mathcal{N}(x; m_n, P_n)\mathcal{N}(x; m_p, P_p) = \mathcal{N}(m_n; m_p, P_p + P_n)\mathcal{N}(x; m_{n|p}, P_{n|p}) \tag{5.68}$$

where

$$m_{n|p} = (P_n^{-1} + P_p^{-1})^{-1}(P_n^{-1}m_n + P_p^{-1}m_p) \tag{5.69}$$

$$P_{n|p} = (P_n^{-1} + P_p^{-1})^{-1} \tag{5.70}$$

Inserting (5.68) to (5.67) and then inserting (5.67) to (5.61), the distance between two Gaussian Mixtures belonging to different time instants are found as in (5.71)

$$D(k_1, k_2) = \sum_{n=1}^{N(k_1)+N(k_2)} \sum_{p=1}^{N(k_1)+N(k_2)} w_n w_p \mathcal{N}(m_n; m_p, P_p + P_n) \tag{5.71}$$

5.7.2.5 Optimal SubPattern Assignment (OSPA) Metric

OSPA distance is proposed by [59] and serves as a metric providing information regarding the difference between two finite sets. Input for the OSPA function are generally taken as the state estimates set of the PHD filter and true target state set. This metric takes both the difference between the state positions and the difference in the set cardinalities into account.

Properties of OSPA provided in [59] are briefly listed in as below:

- a metric on the space of finite sets,
- a natural (meaningful) physical interpretation,
- taking the cardinality errors and state errors into account meaningfully,
- easy computation.

OSPA metric is widely used for performance evaluation of the multi-target state estimators. In the case of this thesis, there does not exist the ground

true target set yet this distance is utilized for finding the difference between given two state estimate set belonging to different time instants of the video, $\widehat{X}(k_1) = \{x_1(k_1), \ldots, x_m(k_1)\}$ and $\widehat{X}(k_2) = \{x_1(k_2), \ldots, x_n(k_2)\}$, where $\widehat{X}$ is the estimated feature position set and $m, n \in \mathbb{N}_0 = \{0, 1, 2, \ldots\}$. The equation of the metric of order p with cut-off c is proposed by [59] as in (5.72).

$$\overline{d}_p^{(c)}(\widehat{X}(k_1), \widehat{X}(k_2)) := \left(\frac{1}{n} \left(\min_{\pi \in \Pi_n} \sum_{i=1}^{m} d^{(c)}(x_i(k_1), x_{\Pi(i)}(k_2))^p + c^p(n-m) \right) \right) \tag{5.72}$$

For any vector x and y, the parameters in (5.72) are described as:

- $d^c(x, y) := \min(c, d(x, y))$
- $c > 0$: cut-off
- Π_k: the set of permutations on $\{1, 2, \ldots, k\}$ for any $k \in \mathbb{N} = \{1, 2, \ldots\}$
- $d(x, y)$: L_2 norm distance between the vectors

The distance (5.72) is calculated for $m \leq n$ and $1 \leq p < \infty$. In case $m > n$, $\overline{d}_p^{(c)}(\widehat{X}(k_1), \widehat{X}(k_2))$ is found as $\overline{d}_p^{(c)}(\widehat{X}(k_2), \widehat{X}(k_1))$.

Note that in this thesis, the OSPA metric of order 2 with cut-off 10 is chosen for finding the difference between the frames belonging to different time instants. The state estimates are taken as the positional information of the mean values belonging to different modes of the Gaussian Mixture PHD, whose weights are greater than 0.5.

CHAPTER 6

PERFORMANCE EVALUATION OF THE PROPOSED APPROACH

In this chapter, we give performance results of the proposed approach for the videos taken in a controlled environment including KTH Database and custom taken occlusion videos. Firstly, we give detailed information about the controlled databases, KTH and custom taken occlusion videos. Then, we give detail information about the experiments which are performed with different parameters. After explaining the experiments, the performance results of each experiment is given as both correctly recognized video percentages and related confusion matrices in KTH database.

Besides, we analyze the effect of using KLT as an alternative to GMPHD on the tracking phase and provide corresponding recognition performance.

Finally, we compare the proposed approach with a well known algorithm, [54], which is denoted as base algorithm, for custom occlusion scenarios and reveal the recognition performances on custom 6 videos taken for this thesis.

6.1 KTH Database

In this thesis, we used KTH video database as a controlled database for action recognition problem. There are 25 individuals performing 6 type of actions:

- Jogging

- Running
- Walking
- Handwaving
- Handclapping
- Boxing

in 4 different environments:

- d1: outdoors
- d2: outdoors with scale variation
- d3: outdoors with different clothes
- d4: indoors

with a frame size 160×120 and 25 frame/second as given in original paper [60].

Figure 6.1 shows a frame example of each type.

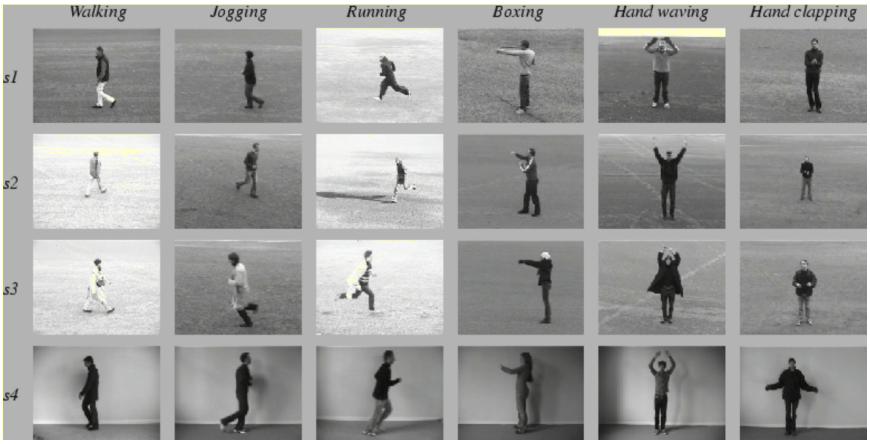


Figure 6.1: Example frames of each video in KTH Database

In this work, we grouped these according to global motion occurrence in the action as given in Table 6.1.

Table 6.1: Defined Groups in the Context of the Thesis

Group Name	Action
Group-1	Jogging, Running, Walking
Group-2	Boxing, Hand Clapping, Hand Waving

Group1 (GR1) actions are the ones with global motion. In these videos, the body in action moves toward one direction. The motion direction of the human body varies as stated below:

1. Motion towards the right direction
2. Motion towards the left direction
3. Diagonal motions towards the lower right and left corners
4. Diagonal motions towards the upper right and left corners

Group-2 actions are the ones without global motion. The person performing the boxing, handclapping, handwaving actions in the video is stationary and the action is caused by the movement of the arms. Videos on the database have specific different characteristics containing:

1. Zoom-in and Zoom-out operation
2. Noisy and noise-free background
3. Different human characteristics (carrying a bag, wearing a coat, male, female, having long hair etc.)
4. Shadow of the human may exist
5. Distance between the person and the camera is different

6.2 Custom Occlusion Videos

The custom video set includes 6 videos in 2 different environment with 3 actions which are walking, running and jogging. The steady objects causing occlusion are chosen as a tree and a person in the videos called Tree-Occluded and Person-Occluded, respectively.

In Figures 6.2 and 6.3, 3 selected frames from the Tree-Occluded and Person-Occluded videos are shown. Notice that the steady object is comparably wide with person being investigated.

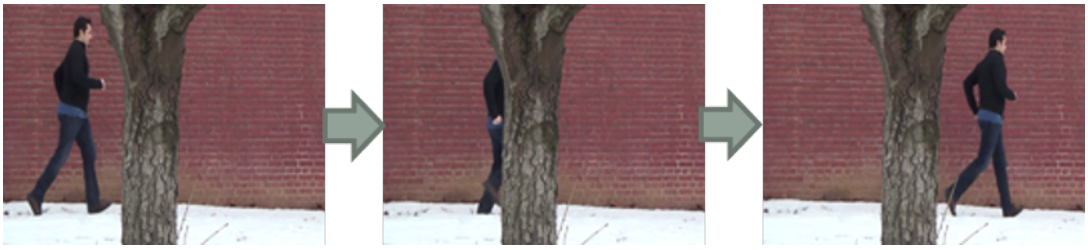


Figure 6.2: Frames from the Tree Occlusion Test Videos

Figure 6.3: Frames from the Person Occlusion Test Videos

Talking about the videos they are taken without using tree-pot, so there is slight motion in the videos. As the environment, they are taken outside and Tree-Occlusion video has more outlier measurements/features which makes recognition difficult.

6.3 Experimental Results on KTH Database

The experiments of the proposed algorithm is performed using the well-known KTH database. And the performance evaluation is done using approximately 50% percent of the videos in KTH database.

6.3.1 Experimental Parameters

Frame Size:

We determine the frame size of the video according the group type. For GR1 actions, the running action determines the frame size since it takes less frame to pass the person in the scene.

Table 6.2: Frame Number for the Groups

Group No	Frame Size
Group-1	14
Group-2	20

Windowing Method:

We implemented 3 different windowing methods as given in Table 6.3.

Table 6.3: Selections of Windowing Type

Windowing Method	Window Size
No windowing	X
Standart Windowing	3
Sliding window	5

In "No windowing" case, we do not perform any windowing operation. Instead, feature vector of each frame is taken and HMM recognition is performed within the information in each frame itself. The **place** of the operation in the flow-chart is given in 6.4.

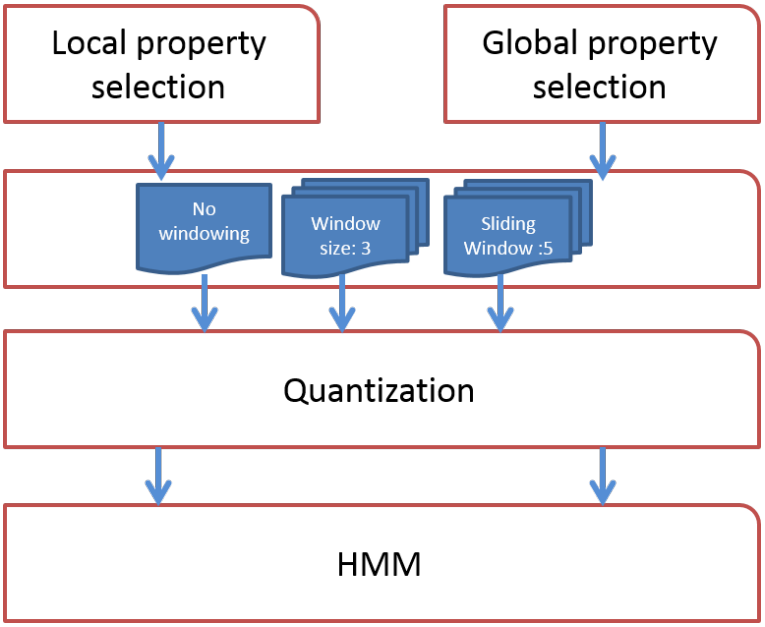


Figure 6.4: Windowing Operations in Flow Chart

In the 2nd windowing approach, we take window size as 3. The feature vector is obtained bu calculating the mean of features for 3 frames. Note that we perform a mean operation in each cluster of each frame. The windowing operation in the FlowChart is given in Figure 6.4.

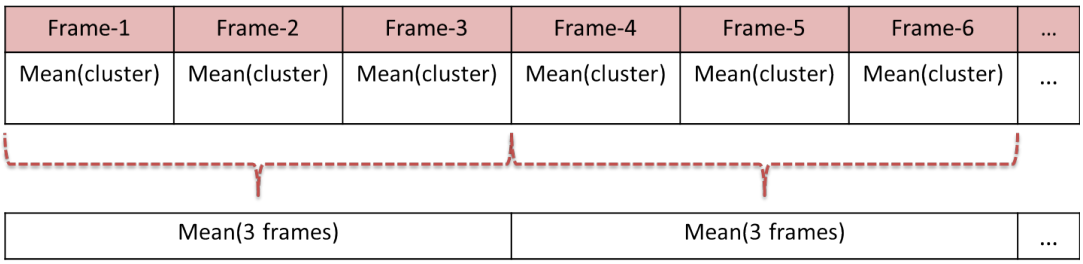


Figure 6.5: Structure of Windowing Method-2

Sliding window approach is utilized as the 3rd windowing operation,6.4. The approach is given in Figure 6.6.

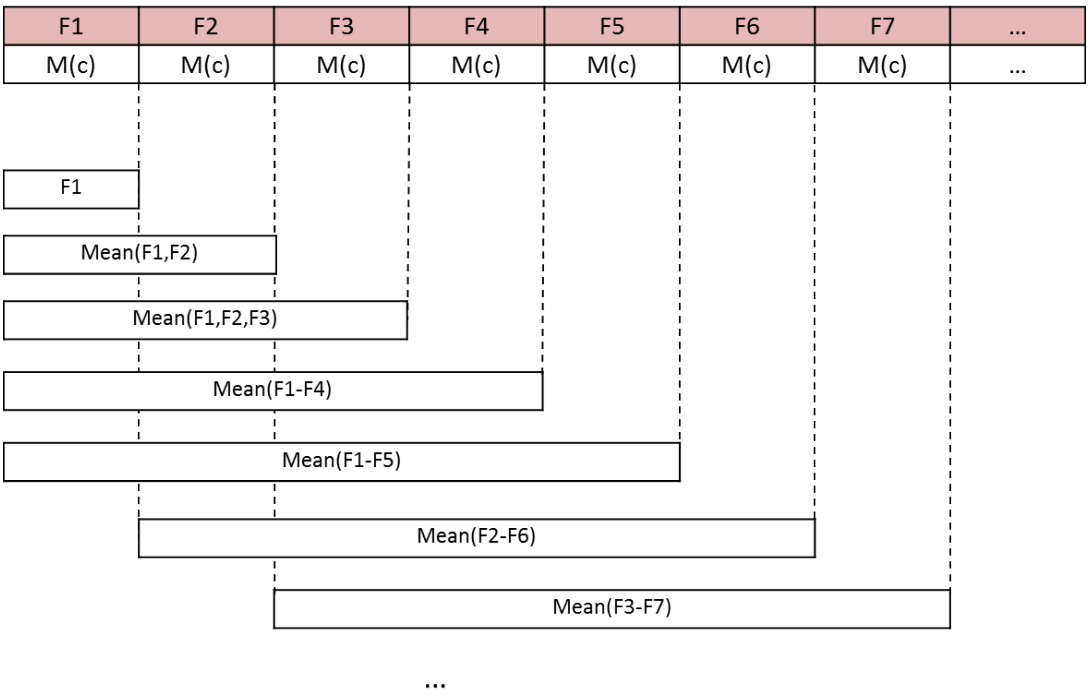


Figure 6.6: Structure of Windowing Method-3

In sling window technique, we obtain feature vector by taking the mean of windows with size 5. In this method, mean operation is performed using the current and previous 4 frame's feature vectors.

HMM Types:

The experiments are performed with using 3 different HMM types to understand the affect of HMM type to recognition performance. HMM Types are given in Table 6.4.

Table 6.4: Selections of HMM Type

HMM Type	Explanation
Type-0	Full HMM
Type-1	Forward HMM
Type-2	Half HMM

Type-0 HMM is ergodic, fully connected one in which we expect transitions

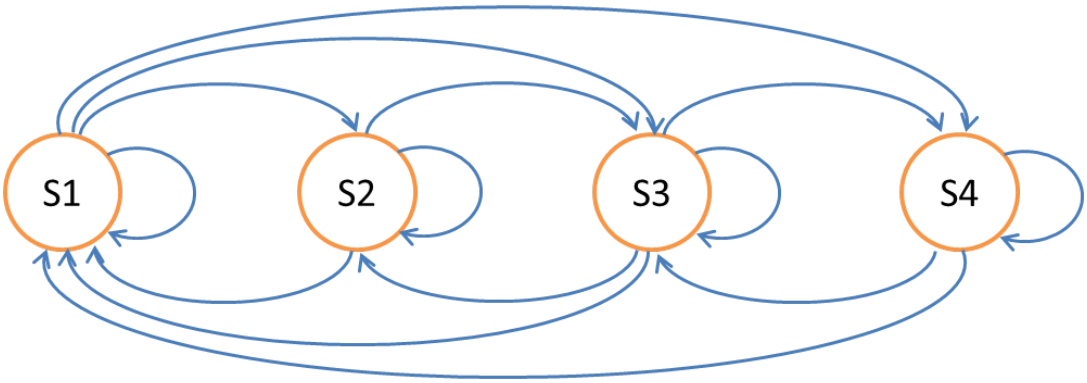


Figure 6.7: Type-0: Fully Connected HMM State Flow

between all hidden states as given in Figure 6.7.

Type-1 HMM is forward connected one in which we expect transitions along forward direction given in Figure 6.8. In action recognition case, when we consider one cycle of action, Type-1 HMM is appropriate.

Figure 6.8: Type-1: Forward Connected HMM State Flow

Type-2 HMM is called half connected HMM in which there is one backward connection to the first hidden state is added compared to Type-1 HMM. This structure is formed considering the existence of cyclic operations in action. The connections are given in Figure 6.9.

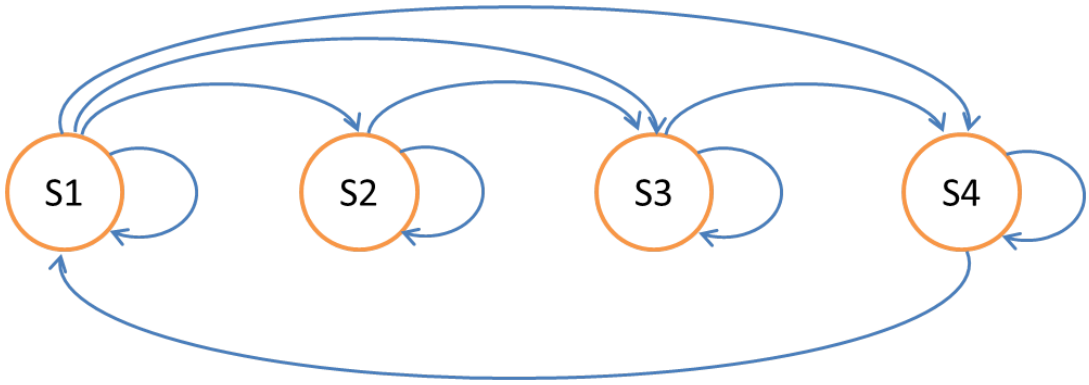


Figure 6.9: Type-2: Half Connected HMM State Flow

6.3.2 Selection of Training Elements

An unique HMM is constructed for each of the feature which is given in Section 5.7 in detail. When we extract these properties of each cluster we obtain 22 and 32 independent HMMs for GR1 and GR2 respectively. The features are listed in the Tables 6.56.6 for GR1 and GR2 action groups respectively.

As mentioned in 5, GR1 has 2 clusters and GR2 has 3 clusters each of which has 10 features. Adding 2 features belonging to body we obtain 22 HMMs for GR1 and 32 HMMs for GR2.

In the experiments, 22 and 32 number of observation sequences are extracted from the video as given in 6.10. Note that Z corresponds to 22 and 32 for GR1 and GR2, respectively.

Table 6.5: Group-1 Training Elements

Element No	Feature Property	Body Segment
1	X Position	Upper Part of the Body
2	X Position	Lower Part of the Body
3	Y Position	Upper Part of the Body
4	Y Position	Lower Part of the Body
5	Direction of Motion	Upper Part of the Body
6	Direction of Motion	Lower Part of the Body
7	Speed	Upper Part of the Body
8	Speed	Lower Part of the Body
9	X Position Std. Dev.	Upper Part of the Body
10	X Position Std. Dev.	Lower Part of the Body
11	Y Position Std. Dev.	Upper Part of the Body
12	Y Position Std. Dev.	Lower Part of the Body
13	Direction of Motion Std. Dev.	Upper Part of the Body
14	Direction of Motion Std. Dev.	Lower Part of the Body
15	Speed Std. Dev.	Upper Part of the Body
16	Speed Std. Dev.	Lower Part of the Body
17	GM Difference	Upper Part of the Body
18	GM Difference	Lower Part of the Body
19	OSPA	Upper Part of the Body
20	OSPA	Lower Part of the Body
21	GM Difference	Whole Body
22	OSPA	Whole Body

Table 6.6: Group-2 Training Elements

Element No	Feature Property	Body Segment
1	X Position	Upper-Left Part of the Body
2	X Position	Upper-Right Part of the Body
3	X Position	Lower Part of the Body
4	Y Position	Upper-Left Part of the Body
5	Y Position	Upper-Right Part of the Body
6	Y Position	Lower Part of the Body
7	Direction of Motion	Upper-Left Part of the Body
8	Direction of Motion	Upper-Right Part of the Body
9	Direction of Motion	Lower Part of the Body
10	Speed	Upper-Left Part of the Body
11	Speed	Upper-Right Part of the Body
12	Speed	Lower Part of the Body
13	X Position Std. Dev.	Upper-Left Part of the Body
14	X Position Std. Dev.	Upper-Right Part of the Body
15	X Position Std. Dev.	Lower Part of the Body
16	Y Position Std. Dev.	Upper-Left Part of the Body
17	Y Position Std. Dev.	Upper-Right Part of the Body
18	Y Position Std. Dev.	Lower Part of the Body
19	Direction of Motion Std. Dev.	Upper-Left Part of the Body
20	Direction of Motion Std. Dev.	Upper-Right Part of the Body
21	Direction of Motion Std. Dev.	Lower Part of the Body
22	Speed Std. Dev.	Upper-Left Part of the Body
23	Speed Std. Dev.	Upper-Right Part of the Body
24	Speed Std. Dev.	Lower Part of the Body
25	GM Difference	Upper-Left Part of the Body
26	GM Difference	Upper-Right Part of the Body
27	GM Difference	Lower Part of the Body
28	OSPA	Upper-Left Part of the Body
29	OSPA	Upper-Right Part of the Body
30	OSPA	Lower Part of the Body
31	GM Difference	Whole Body
32	OSPA	Whole Body

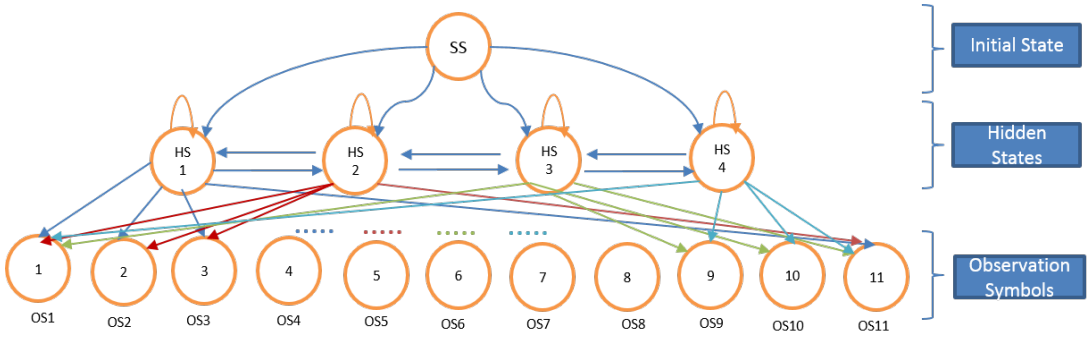


Figure 6.10: An Individual HMM structure

In training phase, all HMM parameters in the HMM structure are trained without making any intelligent selection among the HMMs. In testing phase, again all the recognition results for each video for each HMM is obtained. Then as a post processing optimal selection of features is performed which gives the

6.3.3 Experimented Approaches and Their Performances

The recognition performance of all 9 methods are given in this section. Table 6.7 and Table 6.9gives the performance of the proposed algorithm for each method for GR1and GR2 actions, respectively. The designated results are obtained by optimal selection of features as given in Tables 6.8 and 6.10. After evaluating the output of each HMM and optimal selection is made utilizing these results.

Table 6.7: Methods and Performances for Group-1

Method No	Windowing Type	HMM Type	Perf. for All Videos
1	1	0	89.38%
2	1	1	86.88%
3	1	2	86.88%
4	3	0	85%
5	3	1	86.25%
6	3	2	83.75%
7	5	0	82.5%
8	5	1	83.13%
9	5	2	82.5%

The numbers in Tables 6.8 and Table 6.10 refer to the feature numbers given in Tables 6.5 and 6.6, respectively. The order has the meaning that the ratio of correct recognition increases to the right direction. To speak specially, 19 is number of feature with highest score of the first row of Table 6.8.

Analyzing the results, it is seen that Method1 has the highest recognition performance which is of ergodic HMM type and has no windowing operation.

Similar experiments are performed for Group2(GR2) actions and the results reveals that Method2 has the highest recognition performance for GR2 actions

Table 6.8: Feature Vector Selection for Maximum Throughput for Group-1

Method No	Selected Features for Maximum Performance
1	(1, 3, 5, 6, 7, 8, 10, 11, 13, 14, 15, 16, 19)
2	(1, 3, 4, 5, 6, 7, 8, 11, 13, 14, 15, 16, 19, 22)
3	(1, 3, 5, 7, 8, 9, 10, 11, 14, 15, 16, 19)
4	(1, 5, 7, 10, 11, 12, 13, 15, 16, 19, 20)
5	(3, 6, 7, 8, 10, 12, 14, 15, 16, 19, 20)
6	(1, 2, 3, 6, 7, 15, 16)
7	(1, 7, 8, 12, 13, 14, 15, 19)
8	(1, 3, 7, 8, 10, 15, 16, 20)
9	(2, 3, 7, 14, 15, 16, 20)

which is of forward HMM type and has no windowing operation.

Table 6.9: Methods and Performances for Group-2

Method No	Windowing Type	HMM Type	Perf. for All Videos
1	1	0	86.67%
2	1	1	89.33%
3	1	2	85.33%
4	3	0	87.33%
5	3	1	86%
6	3	2	87.33%
7	5	0	83.33%
8	5	1	87.33%
9	5	2	83.33%

In order to understand and evaluate the recognition performances of the algorithm, confusion matrices give significant information. They show the number and type of misclassified videos.

Confusion Matrices for each experiment are given in Tables 6.11 - 6.19.

Table 6.10: Feature Vector Selection for Maximum Throughput for Group-2

Method No	Selected Features for Maximum Performance
1	(2, 3, 4, 8, 14, 16, 22, 23, 29, 32)
2	(2, 3, 4, 5, 8, 11, 14, 15, 20, 22, 23, 32)
3	(1, 2, 5, 8, 10, 14, 15, 17, 22, 23, 29, 32)
4	(1, 2, 3, 10, 11, 14, 15, 22, 23, 28, 30, 32)
5	(1, 2, 5, 7, 10, 11, 14, 23, 32)
6	(1, 3, 4, 5, 10, 14, 15, 23, 30, 31, 32)
7	(1, 2, 3, 5, 8, 10, 11, 13, 16, 22, 23, 26, 28)
8	(2, 5, 7, 10, 11, 14, 15, 22, 23, 28)
9	(1, 2, 5, 10, 13, 14, 16, 17, 22, 23, 26, 28, 32)

Table 6.11: Confusion Matrix for Windowing Type:1 and HMM Type:0

-	Jg.	Rn.	Wl.	Bx.	Hc.	Hw.
Jg.	45	5	0	0	0	0
Rn.	5	44	0	0	0	0
Wl.	5	2	54	0	0	0
Bx.	0	0	0	42	3	2
Hc.	0	0	0	2	42	6
Hw.	0	0	0	1	6	46

Table 6.12: Confusion Matrix for Windowing Type:3 and HMM Type:0

-	Jg.	Rn.	Wl.	Bx.	Hc.	Hw.
Jg.	42	6	1	0	0	0
Rn.	4	43	2	0	0	0
Wl.	9	2	51	0	0	0
Bx.	0	0	0	39	3	0
Hc.	0	0	0	4	41	3
Hw.	0	0	0	2	7	51

6.3.4 The Effect of OSPA Parameter on the Performance

In this section, the effect of OSPA parameter on the recognition performance is investigated. The 'Feature Set without OSPA' result given in Table 6.20 is obtained by excluding the HMM structures trained by OSPA parameters. For

Table 6.13: Confusion Matrix for Windowing Type:5 and HMM Type:0

-	Jg.	Rn.	Wl.	Bx.	Hc.	Hw.
Jg.	42	6	1	0	0	0
Rn.	4	43	2	0	0	0
Wl.	9	2	51	0	0	0
Bx.	0	0	0	39	3	0
Hc.	0	0	0	4	41	3
Hw.	0	0	0	2	7	51

Table 6.14: Confusion Matrix for Windowing Type:1 and HMM Type:1

-	Jg.	Rn.	Wl.	Bx.	Hc.	Hw.
Jg.	41	5	0	0	0	0
Rn.	7	44	0	0	0	0
Wl.	7	2	54	0	0	0
Bx.	0	0	0	40	5	2
Hc.	0	0	0	2	44	2
Hw.	0	0	0	3	2	50

Table 6.15: Confusion Matrix for Windowing Type:3 and HMM Type:1

-	Jg.	Rn.	Wl.	Bx.	Hc.	Hw.
Jg.	42	7	0	0	0	0
Rn.	4	42	0	0	0	0
Wl.	9	2	54	0	0	0
Bx.	0	0	0	38	4	2
Hc.	0	0	0	4	42	3
Hw.	0	0	0	3	5	49

Group 1 actions we exclude OSPA for upper,lower and whole body parts, for Group 2 actions we exclude OSPA for upper-left, upper-right,lower and whole body parts.

Discussion When we analyze the results, we see that OSPA parameter makes a 3% improvement on recognition performance for Group2 actions. On the other hand it makes no improvement for Group1 actions. From these results we can conclude that the OSPA parameter is not a fundamental but auxiliary parameter

Table 6.16: Confusion Matrix for Windowing Type:5 and HMM Type:1

-	Jg.	Rn.	Wl.	Bx.	Hc.	Hw.
Jg.	41	6	3	0	0	0
Rn.	7	41	0	0	0	0
Wl.	7	4	51	0	0	0
Bx.	0	0	0	41	5	2
Hc.	0	0	0	2	41	3
Hw.	0	0	0	2	5	49

Table 6.17: Confusion Matrix for Windowing Type:1 and HMM Type:2

-	Jg.	Rn.	Wl.	Bx.	Hc.	Hw.
Jg.	41	6	0	0	0	0
Rn.	6	44	0	0	0	0
Wl.	8	1	54	0	0	0
Bx.	0	0	0	40	7	2
Hc.	0	0	0	4	41	5
Hw.	0	0	0	1	3	47

Table 6.18: Confusion Matrix for Windowing Type:3 and HMM Type:2

-	Jg.	Rn.	Wl.	Bx.	Hc.	Hw.
Jg.	34	3	2	0	0	0
Rn.	11	48	0	0	0	0
Wl.	10	0	52	0	0	0
Bx.	0	0	0	42	3	2
Hc.	0	0	0	2	44	7
Hw.	0	0	0	1	4	45

for action recognition.

6.3.5 Discussions

In this chapter, we give the performance results of the proposed algorithm in 9 different experiments with 3 different HMM model and 3 different windowing techniques.

Table 6.19: Confusion Matrix for Windowing Type:5 and HMM Type:2

-	Jg.	Rn.	Wl.	Bx.	Hc.	Hw.
Jg.	39	6	2	0	0	0
Rn.	8	42	1	0	0	0
Wl.	8	3	51	0	0	0
Bx.	0	0	0	37	4	3
Hc.	0	0	0	3	45	8
Hw.	0	0	0	5	2	43

Table 6.20: GMPHD Results for the Feature Sets Including and not Including OSPA Distance

	GR1	GR2
Feature Set with OSPA	89.38%	89.33%
Feature Set without OSPA	89.38%	86.00%

When we analyze the results, the highest performance is achieved for the ergodic HMM for GR1. This is meaningful since the observation data may not contain one cycle of information. So the connection between all states have been necessary. On the other hand, GR2 has the greatest performance with HMM type-1 since the size of the data is long enough to complete one cycle of the action.

Comparing the window size results, the best is achieved in no windowing case for both GR1 and GR2 videos. The reason for this is the reduction in the number of sequence in the observation space for window size 3.Since the observation sequences are not long enough, windowing processes lead loss of data, and the results obtained with the complete data give the best results. For sliding-window technique, i.e. window size is 5, the low-pass filtering (mean operation) results in the reduction of informative data between the frames and when we analyze data in a video, we see that sliding window approach brings the information in the frame to almost same level.

Talking about the selection of features, speed brings out the most informative feature among all for all videos. Speed as well as its standard deviation carries critical information belonging to the action. In addition to the speed, X position,

Y position, direction of upper part and body OSPA differences (for GR2) are discriminative features as given in first row of 6.8 and $2nd$ row of 6.10.

Note that OSPA metric obtained by considering the whole body positions plays a critical role in action recognition of GR2 actions.

We see that GM-PHD distance parameter is not a consistent parameters when we consider the distance between the consecutive frames. It is observed that this inconsistency is caused by the sensitivity of the distance to the covariance information, which can take broad range of values for different features and different time frames.

Another important observation regarding GMPHD filter is that Harris corner detector is not capable of extracting consistent observation points in the videos in consideration due to illumination changes. In some frames Harris may not find some critical points. In these cases PHD is able to compensate the low performance of Harris. PHD can track the observations although the Harris corner Detector does not extract the observation during a few (4-5) frames.

Talking about the clustering phase, we divide GR-1 action into two and GR-3 action into three. The reason for this is the moment operation we apply to the features. For GR-2 actions, the upper part action for handwaving and handclapping is symmetric. So the change in the mean of the velocities and positions do not observed. So we divide the upper part into two and we can observe the characteristics of the activity in moments of action and able to make recognition.

6.3.6 Discussion of the Algorithm for Misclassified Videos

When we analyze misclassified videos, there are two major effects prominent at first glance. One of them is the characteristics of the video such that the actions which are performed slower and faster than the nominal speeds are misclassified. An example to this situation is classification of jogging videos as walking or running, or vice verse. The direction of action is also very critical when we analyze the videos in which the human moves diagonally or towards the camera.

Although we take cautions to eliminate this effect by resizing the body, the speed and directions of legs and foot are become non-perceptible by the tracker which results in reduction in the speed of lower part and cause misclassification.

On the other hand, since speed and direction of motion of lower part is critical for the recognition of GR1, extraction observations belonging to foot and accurate tracking of these plays an important role. So, when they are not extracted or eliminated in descriptive mode selection phases, it directly affects the result and cause misclassification.

Considering GR2 actions, we see that carriage of bag and topcoat affects the recognition performance. Beside this, one of the major drawbacks comes from symmetry operation in which the center of lower part is extracted and utilized for upper part clustering. Although we eliminate the outliers of lower part with position information, there might left non eliminated outliers which cause antisymmetric clustering and misclassification. Note that if the extracted observations are not symmetric (no observation on one of the arms), then the action is not recognized correctly.

GR2 action is also affected from zooming operation since the addition of y velocity component. Analyzing the misclassified videos of hand-clapping as boxing reveals that there is similarity between these action when just one of the upper parts considered. When we focus on clapping of one hand and boxing, the action properties become close in terms of the x velocity and position. So utilization of upper part information of GR2 actions as a whole would increase the recognition performance.

Note that GM-PHD is capable of eliminating noise in the frames by its inner mechanism successfully. It does not include the Harris observation, which are not persistent and away from the body, to the group tracking.

6.4 Comparison between KLT and GMPHD

In this section the experimental results of the proposed approach but KLT tracking are examined. The usage of KLT can be seen in Figure 4.7. In the observation extraction step, Harris Corner Detector is utilized and the observations are extracted in the very first frame. Then they are fed to the KLT tracker which gives the location of the points in the next frame as output. As in the proposed approach, the action directions are rotated, HLR is performed and body clustering is made. Different from GMPHD, the velocity information is extracted from the difference in positions in KLT. Another difference is in the number of HMM structures. Since there is no GM-PHD difference in KLT, we utilize 3 and 4 less number of HMM structure for Group1 and Group2 respectively.

The KLT parameters are selected in such a way that it allows high displacement for even the running action. The number of pyramid levels are taken as 3 where image pyramids are constructed by decreasing the resolution of the image in half for each level. By using different levels, the points are tracked in different resolution. All the extracted points are tracked in each frame and the most similar point is found.The block size is taken as 31×31 which is the pixel size of the patch used for spatial gradient computation around each tracked point. If the KLT tracker can not find the corresponding match, the algorithm renews/refinds the observations.

The experiments are performed exactly in the same way using the same parameter sets and the video sets. Similar to proposed approach, the parameters are selected optimally to obtain best score.

The performance using KLT is given in Table 6.21. When we compare the results with the proposed algorithm given in Table 6.20, we see there is about 4% performance degradation when we use KLT as the tracker instead of GM-PHD.

Talking about the effect of OSPA parameter to recognition performance, OSPA brings more advantage for KLT than GMPHD for GR1 actions. Note that the number of observations kept almost constant in KLT tracker. So we can com-

Table 6.21: KLT Results for the Feature Sets Including and not including OSPA Distance

	GR1	GR2
Feature Set with OSPA	85.625%	86.87%
Feature Set without OSPA	82.5%	83.125%

ment on the result as the change in position information between consecutive frames effect the overall performance than the change in the number of observations does.

Discussions

In this section, the KLT track performance are examined and comparison between tracking/group tracking performance between KLT and PHD is given in terms of action recognition capability. It is observed that the performance of proposed algorithm is higher that of the one using KLT as a tracker.

The reasons for this performance difference can be listed as below:

1. In KLT, the observations are extracted in the very first frame and the next position of the point is tracked/found in the next frame. And when the observation is lost the observations are re-extracted similar to GM-PHD. The number of features are changed in case of lost.

2. GMPHD gives a complete solution including the velocity estimates. On the other hand, in KLT we calculate the velocities by position difference as a post process.

Tracking performance of KLT is robust when the block size is adjusted according to the action. However, when there is occlusion, KLT misses the occluded observations. But since the observations are re-extracted, the information about the region centered the observation point is changed which may cause reduction in recognition performance.

6.5 Comparison with Literature

For benchmarking purposes, we selected an algorithm in the literature which has high recognition performance on KTH Database. The videos on KTH Database do not contain occlusion scenarios. So in order to reveal the performances of the base and proposed algorithms on occluded scenarios, we obtain/took videos and the results are given in this section. in [54] as base algorithm and make comparison with the proposed one on custom taken occlusion videos. we denote the method using tracklet descriptor as base algorithm from now on. The reason of selecting [54] as base algorithm is that it has high, 94.5%, recognition performance on KTH database.

The aim is to compare recognition performances of both algorithm in the scenarios containing occlusion.

Another important step for evaluating the performance of the proposed algorithm is to compare it with an existing algorithm on custom videos. The existing approach is selected such that it has higher performance on the same database. The idea here is to reveal the power of GMPHD filter in occlusion scenarios by comparing it with the existing approach.

To enable this, we took custom videos including occlusion scenarios and used these videos for benchmarking.

Compared Algorithm: Tracklet

We selected [54] as base algorithm in the literature due to its high performance on KTH database. In this section we briefly talk about its approach.

The algorithm extracts KLT features as observations and tracks these with KLT tracker. It replaces the lost ones with the new as the observations are lost. Then tracklet descriptors are constructed by calculating local appearance (HOG) and local motion (HOF) features around the tracked points at each frame constructing varying length feature time series. Using bag of features technique, the algorithm maps each trajectory in a video to the closest word(cluster) which

allows to represent each video by histogram of occurrences. In dictionary construction phase, Dynamic Time Warping distance is utilized and the number of clusters are calculated by Affinity Propagation. Note that in dictionary construction phase we use the same videos (3 person, 3 action 4 environment making up 36) in KTH database as proposed in the paper. In the recognition phase, the algorithm uses SVM with x^2 radial basis function for kernel. In order to train the SVM, we use the same videos in proposed approach. And the tests are performed with the taken occluded videos.

Summary of Base and Proposed Algorithms

Image Features: The image feature extraction techniques are similar in both techniques.Note that these are denoted as observation/measurement. Tracking: Based algorithm uses intensity based tracking approach, KLT, where proposed one uses statistical group tracking which does not uses intensity values directly, Descriptor: Based algorithms uses tracklet descriptors which combines time and space information by extracting HOG anf HOF features. Proposed algorithm uses GMPHD filter outputs and their moments, which does not include time but space information, obtained in each frame. Recognition: Based algorithm uses SVM classification technique. Proposed algorithm uses HMM for recognition. Since tracklet descriptor of base algorithm includes time information in its structure, SVM classification technique is adequate for recognition. On the other hand, proposed approach uses HMM since it includes the relation between sequence elements. But similarly, SVM classification method may be used by concatenating the features.

Experimental Results

Both algorithms are trained with the same data subset in KTH database and they are tested with the occlusion videos. The resultant performances are provided in Table 6.22.

Performance Evaluation

Tracklet approach has 67% recognition performance which shows it is capable of handling occlusion cases. The approach eliminates short paths and makes recog-

Table 6.22: Performances of Tracklet and the Proposed Methods for the Occlusion Scenarios ('W' stands for Walking action)

Object Causing Occlusion	Action	Tracklet Method	Proposed Method
Tree	Walking	✓	✓
	Running	✗(W)	✓
	Jogging	✗(W)	✗(W)
Person	Walking	✓	✓
	Running	✓	✓
	Jogging	✓	✓

nition using histogram property which does not make significant performance degradation even if it can not track the body when it is behind the object. The algorithm re-finds the missed observations in each frame which increases the recognition performance.

Our proposed approach is capable of correctly recognizing 83% of the custom videos. After analyzing the videos misclassified by base algorithm but ours, we see that the velocity estimation of the GMPHD is more reliable than the optical flow in this scenario. Our approach eliminates inconsistent objects in GMPHD phase and eliminates slow object as a post processing before recognition which increases it performance.We use moments of the position and velocity information which help eliminate noise in the video. Additionally, it can re-track the spawn measurements by means of birth mechanism in a very short time instant, mostly at the time of appearance.

The performance of both algorithms have 100% performance in the Person-Occlusion video since it is a relatively easy scenario in terms of noise. In Tree-Occlusion video, there are slowly moving small objects which decreases the overall velocity and changes the shape of the tracked object. And this causes mis-classification.

Talking about computational cost, base algorithm has high computational load on tracklet descriptor extraction phase since HOF features are extracted on SIFT. On the other hand, classification phase has a very low computational cost since it uses SVM. Similarly, proposed algorithm has comparable computational

cost on GMPHD filtering. The training phase of HMM is more costly than SVM, but testing phases of both algorithms take just a few seconds.

6.6 Discussion

In this chapter we reveal the performance of the proposed algorithm on KTH database and make a comparison between a well known recognition approach. Besides the (group) tracking performance of the algorithm is compared with the KLT.

When we analyze the results the proposed approach is capable of handling the occlusion which shows the power of GMPHD in these scenarios due to its birth and update mechanisms.

CHAPTER 7

CONCLUSION

This thesis addresses Human Action Recognition problem and proposes a complete solution enabling the recognition of human actions in the video sequences. The proposed solution includes all the necessary steps, like detection, tracking and recognition and utilizes GMPHD which is a state-of-the-art multi-target tracking approach showing that GMPHD is applicable to human action recognition problem. Detection is performed by applying HCD technique to individual image frames and the detection outputs (HCD observations) are taken as independent measurements. This dynamically changing measurement set is fed to GMPHD filter and intensity function is obtained. Estimated positions and other relevant information, like speed, velocity, standard deviation etc., regarding the measurements are extracted and transformed into the appropriate format. Estimations belonging to different body parts are also considered as descriptive information. Finally, a structure including several paralel HMM's is used for both training and testing phases and recognition of the actions is completed.

The experiments reveal that the technique can recognize 89.3% of the videos in KTH Database correctly. This result shows that the performance of the approach is comparable with those of existing techniques in the literature.

To understand the drawbacks of the method, the remaining 10% unrecognized videos are investigated so as to describe the reasons for the failure.

Two types of benchmarking is performed to evaluate the performance of the proposed algorithm. Firstly, the GMPHD group tracking filter in the proposed

approach is replaced with KLT tracker and recognition performances are compared. Results show that GMPHD has about 10% higher performance than the one with KLT. Secondly, the occlusion responses of the proposed and existing algorithms are tested with custom 6 videos. Both existing and proposed algorithms are able to handle occlusion with one more positive recognition of proposed one.

Novelties and contributions of the thesis are mainly around the usage of GMPHD filter for solving human action recognition problem and can be briefly listed as:

- Performing group tracking of the features by the GMPHD filter for solving human action recognition problem
- Extraction of information in the shape of a 3D GMPHD function derived from 2D video frames
- Utilization of OSPA and GMPHD differences between the frames belonging to different time instants together with the information regarding the estimated feature states
- Inclusion of image intensity difference as a function into the GMPHD measurement update stage

In summary, this thesis raises a novel approach to deal with human action recognition problem and presents its performance for a well-known video data base. In this work, we made an initial study on using GMPHD filter for action recognition and realize that it can be used for recognizing human actions. Definitely, there are still more works to be done to reveal the power of GMPHD and to increase its performance which can be:

- Using GMPHD in multiple object scenarios. The power of PHD is in multi-target group tracking and it can be used as an action recognizer in scenarios having multiple objects.
- Using different velocity motion models in GMPHD which fits to the action being performed, i.e. sinusoidal motions for GR1 type actions, and analyze the performance of the approach.

- Eliminate outliers in GMPHD Filtering phase. Constant velocity features can be ignored if we want to extract motion information.
- Performing priority weighting for the descriptive feature parameters during the descriptive feature selection phase.
- Utilizing different feature extraction techniques.
- Analyzing the performance effect of OSPA and GMPHD distances between second or more order differences in time (i.e., not only for consecutive frames).

REFERENCES

[1] J. Aggarwal and M. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43(3):16:1–16:43, Apr. 2011.

[2] J. G. Allen, R. Y. Xu, and J. S. Jin. Object tracking using camshift algorithm and multiple quantized feature spaces. In *Proceedings of the Pan-Sydney area workshop on Visual information processing*, pages 3–7. Australian Computer Society, Inc., 2004.

[3] L. Antón-Canalís, E. Sánchez-Nielsen, and M. Hernández-Tejera. Swarmtrack: A particle swarm approach to visual tracking. In *Proceedings of the international Conference on Computer vision theory and applications*, volume 2, pages 221–228, 2006.

[4] M. J. Black and A. D. Jepson. Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision*, 26(1):63–84, 1998.

[5] A. F. Bobick and A. D. Wilson. A state-based approach to the representation and recognition of gesture. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(12):1325–1337, 1997.

[6] M. Bregonzio, S. Gong, and T. Xiang. Recognising action as clouds of space-time interest points. *CVPR*, 2009.

[7] T. J. Broida and R. Chellappa. Estimation of object motion parameters from noisy images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (1):90–99, 1986.

[8] H. H. Bui, D. Q. Phung, and S. Venkatesh. Hierarchical hidden markov models with general state hierarchy. In *Proceedings of the National Conference on Artificial Intelligence*, pages 324–329. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004.

[9] H. H. Bui, S. Venkatesh, and G. West. Policy recognition in the abstract hidden markov model. *Journal of Artificial Intelligence Research*, pages 451–499, 2002.

[10] T.-J. Cham and J. M. Rehg. A multiple hypothesis approach to figure tracking. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 2. IEEE, 1999.

[11] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(5):564–577, 2003.

[12] T. Darrell and A. Pentland. Space-time gestures. In *Computer Vision and Pattern Recognition, 1993. Proceedings CVPR'93., 1993 IEEE Computer Society Conference on*, pages 335–340. IEEE, 1993.

[13] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Proceedings of the 14th International Conference on Computer Communications and Networks*, ICCCN '05, pages 65–72, Washington, DC, USA, 2005. IEEE Computer Society.

[14] T. V. Duong, H. H. Bui, D. Q. Phung, and S. Venkatesh. Activity recognition and abnormality detection with the switching hidden semi-markov model. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 838–845. IEEE, 2005.

[15] H. A. A. El-Halym, I. I. Mahmoud, A. AbdelTawab, and S. Habib. Particle filter versus particle swarm optimization for object tracking. In *Proceedings of ASAT Conf*, 2009.

[16] A. Gilbert, J. Illingworth, and R. Bowden. Fast realistic multi-action recognition using mined dense spatio-temporal features. In *ICCV*, pages 925–931. IEEE, 2009.

[17] I. R. Goodman, R. P. Mahler, and H. T. Nguyen. *Mathematics of data fusion*, volume 37. Springer Science & Business Media, 2013.

[18] J. Han and B. Bhanu. Individual recognition using gait energy image. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(2):316–322, 2006.

[19] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, page 50. Citeseer, 1988.

[20] S. Hernandez and M. Frean. Bayesian multiple person tracking using probability hypothesis density smoothing. *International Journal on Smart Sensing and Intelligent Systems*, 4(2):285–312, 2011.

[21] S. Hongeng and R. Nevatia. Large-scale event detection using semi-hidden markov models. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1455–1462. IEEE, 2003.

[22] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. Huang. Action Detection in Complex Scenes with Spatial and Temporal Ambiguities inProceedings. pages 128–135, 2009.

[23] C. Hue, J.-P. Le Cadre, and P. Pérez. Sequential monte carlo methods for multiple target tracking and data fusion. *Signal Processing, IEEE Transactions on*, 50(2):309–325, 2002.

[24] N. Ikizler-Cinbis and S. Sclaroff. Object, scene and actions: Combining multiple features for human action recognition. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *ECCV (1)*, volume 6311 of *Lecture Notes in Computer Science*, pages 494–507. Springer, 2010.

[25] M. Isard and A. Blake. Condensation—conditional density propagation for visual tracking. *International journal of computer vision*, 29(1):5–28, 1998.

[26] M. Isard and J. MacCormick. Bramble: A bayesian multiple-blob tracker. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 34–41. IEEE, 2001.

[27] A. D. Jepson, D. J. Fleet, and T. F. El-Maraghi. Robust online appearance models for visual tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(10):1296–1311, 2003.

[28] S. Jones, L. Shao, J. Zhang, and Y. Liu. Relevance feedback for real-world human action retrieval. *Pattern Recognition Letters*, pages 446–452, 2012.

[29] V. Kellokumpu, G. Zhao, and M. Pietikäinen. Recognition of human actions using texture descriptors. *Machine Vision and Applications*, 22(5):767–780, 2011.

[30] J. Kennedy. Particle swarm optimization. In *Encyclopedia of Machine Learning*, pages 760–766. Springer, 2010.

[31] Z. Khan, T. Balch, and F. Dellaert. Mcmc-based particle filtering for tracking a variable number of interacting targets. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(11):1805–1819, 2005.

[32] T.-K. Kim and R. Cipolla. Canonical correlation analysis of video volume tensors for action categorization and detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(8):1415–1428, Aug. 2009.

[33] W. Kim, J. Lee, M. Kim, D. Oh, and C. Kim. Human action recognition using ordinal measure of accumulated motion. *EURASIP J. Adv. Signal Process*, 2010:2:1–2:10, Feb. 2010.

[34] G. Kitagawa. Non-gaussian state—space modeling of nonstationary time series. *Journal of the American statistical association*, 82(400):1032–1041, 1987.

[35] I. Laptev. On space-time interest points. *Int. J. Comput. Vision*, 64(2-3):107–123, Sept. 2005.

[36] Z. Lin, Z. Jiang, and L. S. Davis. Recognizing actions by shape-motion prototype trees. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 444–451. IEEE, 2009.

[37] C. Liu and P. C. Yuen. Human action recognition using boosted eigenactions. *Image Vision Comput.*, 28(5):825–835, 2010.

[38] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'81, pages 674–679, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc.

[39] Y. M. Lui and J. R. Beveridge. Tangent bundle for human action recognition. In *FG*, pages 97–102. IEEE, 2011.

[40] F. Lv and R. Nevatia. Single view human action recognition using key pose matching and viterbi path searching. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[41] J. MacCormick and A. Blake. A probabilistic exclusion principle for tracking multiple objects. *International Journal of Computer Vision*, 39(1):57–71, 2000.

[42] D. J. MacKay. Introduction to monte carlo methods. In *Learning in graphical models*, pages 175–204. Springer, 1998.

[43] E. Maggio, E. Piccardo, C. Regazzoni, and A. Cavallaro. Particle phd filtering for multi-target visual tracking. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 1, pages I–1101. IEEE, 2007.

[44] R. Mahler. *Statistical Multisource-Multitarget Information Fusion*. Artech House, Norwood, MA, 2007.

[45] R. P. Mahler. Multitarget bayes filtering via first-order multitarget moments. *Aerospace and Electronic Systems, IEEE Transactions on*, 39(4):1152–1178, 2003.

[46] R. Messing, C. Pal, and H. A. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *ICCV*, pages 104–111. IEEE, 2009.

[47] D. Musicki, R. Evans, and S. Stankovic. Integrated probabilistic data association. *Automatic Control, IEEE Transactions on*, 39(6):1237–1241, 1994.

[48] P. Natarajan and R. Nevatia. Coupled hidden semi markov models for activity recognition. In *Motion and Video Computing, 2007. WMVC'07. IEEE Workshop on*, pages 10–10. IEEE, 2007.

[49] H. T. Nguyen and B. Bhanu. Multi-object tracking in non-stationary video using bacterial foraging swarms. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 877–880. IEEE, 2009.

[50] K. Okuma, A. Taleghani, N. De Freitas, J. J. Little, and D. G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *Computer Vision-ECCV 2004*, pages 28–39. Springer, 2004.

[51] E. Pollard, A. Plyer, B. Pannetier, F. Champagnat, and G. L. Besnerais. Gm-phd filters for multi-object tracking in uncalibrated aerial videos. In *Information Fusion, 2009. FUSION'09. 12th International Conference on*, pages 1171–1178. IEEE, 2009.

[52] R. Poppe. A survey on vision-based human action recognition. *Image Vision Comput.*, 28(6):976–990, June 2010.

[53] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[54] M. Raptis and S. Soatto. Tracklet descriptors for action modeling and video analysis. In *Proceedings of the 11th European Conference on Computer Vision: Part I*, ECCV'10, pages 577–590, Berlin, Heidelberg, 2010. Springer-Verlag.

[55] C. Rasmussen and G. D. Hager. Probabilistic data association methods for tracking complex visual objects. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(6):560–576, 2001.

[56] D. B. Reid. An algorithm for tracking multiple targets. *Automatic Control, IEEE Transactions on*, 24(6):843–854, 1979.

[57] J. M. D. Rincón, D. Makris, C. O. Uruňuela, and J.-C. Nebel. Tracking human position and lower body parts using kalman and particle filters constrained by human biomechanics. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 41(1):26–37, 2011.

[58] S. Sadek, A. Al-Hamadi, B. Michaelis, and U. Sayed. An action recognition scheme using fuzzy log-polar histogram and temporal self-similarity. *EURASIP J. Adv. Sig. Proc.*, 2011, 2011.

[59] D. Schuhmacher, B.-T. Vo, and B.-N. Vo. A consistent metric for performance evaluation of multi-object filters. *IEEE Transactions on Signal Processing*, 56(8):3447–3457, 2008.

[60] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. pages 32–36, 2004.

[61] Q. Shi, L. Cheng, L. Wang, and A. Smola. Human action segmentation and recognition using discriminative semi-markov models. *International journal of computer vision*, 93(1):22–32, 2011.

[62] R. L. Streit and T. E. Luginbuhl. Maximum likelihood method for probabilistic multihypothesis tracking. In *SPIE's International Symposium on Optical Engineering and Photonics in Aerospace Sensing*, pages 394–405. International Society for Optics and Photonics, 1994.

[63] T. H. Thi, J. Zhang, L. Cheng, L. Wang, and S. Satoh. Human action recognition and localization in video using structured learning of local space-time features. In *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pages 204–211. IEEE, 2010.

[64] M. Z. Uddin, J. Lee, and T.-S. Kim. Independent shape component-based human activity recognition via hidden markov model. *Applied Intelligence*, 33(2):193–206, 2010.

[65] A. Veeraraghavan, R. Chellappa, and A. K. Roy-Chowdhury. The function space of an activity. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 959–968. IEEE, 2006.

[66] R. Vezzani, D. Baltieri, and R. Cucchiara. Hmm based action recognition with projection histogram features. In *Proceedings of the 20th International Conference on Recognizing Patterns in Signals, Speech, Images, and Videos*, ICPR'10, pages 286–293, Berlin, Heidelberg, 2010. Springer-Verlag.

[67] B.-N. Vo and W.-K. Ma. The Gaussian mixture probability hypothesis density filter. *IEEE Transactions on Signal Processing*, 54(11):4091–4104, 2006.

[68] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 3169–3176, Washington, DC, USA, 2011. IEEE Computer Society.

[69] Q. Wang, L. Xie, J. Liu, and Z. Xiang. Enhancing particle swarm optimization based particle filter tracker. In *Computational Intelligence*, pages 1216–1221. Springer, 2006.

[70] Y.-D. Wang, J.-K. Wu, W. Huang, A. Kassim, et al. Gaussian mixture probability hypothesis density for visual people racking. In *Information Fusion, 2007 10th International Conference on*, pages 1–6. IEEE, 2007.

[71] Y.-D. Wang, J.-K. Wu, A. Kassim, W.-M. Huang, et al. Tracking a variable number of human groups in video using probability hypothesis density. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 1127–1130. IEEE, 2006.

[72] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on*, pages 379–385. IEEE, 1992.

[73] E. Yu and J. Aggarwal. Human action recognition with extremities as semantic posture representation. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2009.

[74] X. Zhang, W. Hu, S. Maybank, X. Li, and M. Zhu. Sequential particle swarm optimization for visual tracking. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[75] X. Zhang, W. Hu, W. Qu, and S. Maybank. Multiple object tracking via species-based particle swarm optimization. *Circuits and Systems for Video Technology, IEEE Transactions on*, 20(11):1590–1602, 2010.

[76] M. Ziaeefard and H. Ebrahimnezhad. Hierarchical human action recognition by normalized-polar histogram. In *ICPR*, pages 3720–3723. IEEE Computer Society, 2010.

APPENDIX A

APPENDIX CHAPTER

Table A.1: GMPHD filter (Prediction of birth targets, prediction of existing targets, construction of PHD update components steps),(adopted from [67]).

given $\left\{w_{k-1}^{(i)}, m_{k-1}^{(i)}, P_{k-1}^{(i)}\right\}_{i=1}^{J_{k-1}}$ and the measurement set Z_k

step 1. (Prediction of birth targets)

$i = 0$

for $j = 1, \ldots, J_{\gamma,k}$

$i := i + 1$

$w_{k|k-1}^{(i)} = w_{\gamma,k}^{(j)},\ m_{k|k-1}^{(i)} = m_{\gamma,k}^{(j)},\ P_{k|k-1}^{(i)} = P_{\gamma,k}^{(j)}$

end

for $j = 1, \ldots, J_{\beta,k}$

for $l = 1, \ldots, J_{k-1}$

$i := i + 1$

$w_{k|k-1}^{(i)} = w_{k-1}^{(l)} w_{\beta,k}^{(j)},\ m_{k|k-1}^{(i)} = d_{\beta,k-1}^{(j)} + F_{\beta,k-1}^{(j)} m_{k-1}^{(l)}$

$P_{k|k-1}^{(i)} = Q_{\beta,k-1}^{(j)} + F_{\beta,k-1}^{(j)} P_{k-1}^{(l)} [F_{\beta,k-1}^{(j)}]^T$

end

end

step 2. (Prediction of existing targets)

for $j = 1, \ldots, J_{k-1}$

$i := i + 1$

$w_{k|k-1}^{(i)} = p_{S,k} w_{k-1}^{(j)}$

$m_{k|k-1}^{(i)} = F_{k-1} m_{k-1}^{(j)}$

$P_{k|k-1}^{(i)} = Q_{k-1} + F_{k-1}^{(j)} P_{k-1}^{(j)} [F_{k-1}^{(j)}]^T$

end

$J_{k|k-1} = i$

step 3. (Construction of PHD update components)

for $j = 1, \ldots, J_{k|k-1}$

$\eta_{k|k-1}^{(j)} = H_k m_{k|k-1}^{(j)}$

$S_k^{(j)} = R_k + H_k P_{k|k-1}^{(j)} [H_k]^T$

$K_k^{(j)} = P_{k|k-1}^{(j)} [H_k]^T [S_k^{(j)}]^{-1}$

$P_{k|k}^{(j)} = [I - K_k^{(j)} H_k] P_{k|k-1}^{(j)}$

end

Table A.2: GMPHD filter (Measurement update and outputting steps), (adopted from [67]).

step 4. (Update)

for $j = 1, \ldots, J_{k|k-1}$

$w_k^{(j)} = (1 - p_{D,k}) w_{k|k-1}^{(j)}$,

$m_k^{(j)} = m_{k|k-1}^{(j)}$, $P_k^{(j)} = P_{k|k-1}^{(j)}$

end

l:=0

for each $z \in Z_k$

l:=l+1

for $j = 1, \ldots, J_{k|k-1}$

$w_k^{(lJ_{k|k-1}+j)} = p_{D,k} w_{k|k-1}^{(j)} \mathcal{N}(z; \eta_{k|k-1}^{(i)}, S_k^{(j)})$

$m_k^{(lJ_{k|k-1}+j)} = m_{k|k-1}^{(j)} + K_k^{(j)}(z - \eta_{k|k-1}^{(j)})$

$P_k^{(lJ_{k|k-1}+j)} = P_k^{(j)}$

end

end

$w_k^{(lJ_{k|k-1}+j)} := \frac{w_k^{(lJ_{k|k-1}+j)}}{\kappa_k(z) + \sum_{i=1}^{J_{k|k-1}} w_k^{(lJ_{k|k-1}+i)}}$, for $j = 1, \ldots, J_{k|k-1}$

$J_k = lJ_{k|k-1} + J_{k|k-1}$

output. $\left\{ w_k^{(i)}, m_k^{(i)}, P_k^{(i)} \right\}_{i=1}^{J_k}$

Table A.3: GMPHD filter (Pruning step), (adopted from [67]).

given $\left\{w_k^{(i)}, m_k^{(i)}, P_k^{(i)}\right\}_{i=1}^{J_k}$, a truncation period T, a merging threshold U, and a maximum allowable number of Gaussian terms J_{max},
Set $l = 0$, and $I = i = 1, \ldots, J_k | w_k^{(i)} > T$.

repeat
- $l := l + 1$
- $j := \arg\max_{i \in I} w_k^{(i)}$
- $L := \left\{i \in I | (m_k^{(i)} - m_k^{(j)})^T (P_k^{(j)})^{-1} (m_k^{(i)} - m_k^{(j)}) \leq U\right\}$
- $\widetilde{w}_k^{(l)} = \sum_{i \in L} w_k^{(i)}$
- $\widetilde{m}_k^{(l)} = \frac{1}{\widetilde{w}_k^{(l)}} \sum_{i \in L} w_k^{(i)} x_k^{(i)}$
- $\widetilde{P}_k^{(l)} = \frac{1}{\widetilde{w}_k^{(l)}} \sum_{i \in L} w_k^{(i)} (P_k^{(i)} + (\widetilde{m}_k^{(l)} - m_k^{(i)})(\widetilde{m}_k^{(l)} - m_k^{(i)})^T)$
- $I := I \backslash L$

until $I = \emptyset$

if $l > J_{max}$ then replace $\left\{\widetilde{w}_k^{(i)}, \widetilde{m}_k^{(i)}, \widetilde{P}_k^{(i)}\right\}_{i=1}^{l}$ by those of the J_{max} Gaussians with largest weights.

output $\left\{\widetilde{w}_k^{(i)}, \widetilde{m}_k^{(i)}, \widetilde{P}_k^{(i)}\right\}_{i=1}^{l}$ as pruned Gaussian components.

Table A.4: GMPHD filter (Multitarget state extraction), (adopted from [67]).

given $\left\{w_k^{(i)}, m_k^{(i)}, P_k^{(i)}\right\}_{i=1}^{J_k}$

Set $\widehat{X}_k = \emptyset$
for $j = 1, \ldots, J_k$
  if $w_k^{(i)} > 0.5$
    for $j = 1, \ldots, round(w_k^{(i)})$
      update $\widehat{X}_k = \left[\widehat{X}_k, m_k^{(i)}\right]$
    end
  end
end

output $\widehat{X}_k$ as the multi-target state estimate.

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Günay, Elif Erdem
Nationality: Turkish (TC)
Date and Place of Birth: 20.10.1982, İzmir
Marital Status: Married
Phone: +90-312-2863731 **Fax:** NA

EDUCATION

Degree	Institution	Year of Graduation
M.S.	METU Electrical and Electronics Engineering	2007
B.S.	METU Electrical and Electronics Engineering	2004
High School	İzmir Science High School	2000

PROFESSIONAL EXPERIENCE

Year	Place	Enrollment
2004-Present	ASELSAN Inc.	Systems Engineer

PUBLICATIONS

National Conference Publications

- E. Erdem, İ. Gürel, H. Yürük, S. Cinel, "Image Processing Applications in Unmanned Systems" (in Turkish), Defense Technologies Conference,

Ankara-Turkey, 2008