

TRACKING HUMAN IN THERMAL VISION USING MULTI-FEATURE
HISTOGRAM

A Thesis
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
Of the Requirements for the Degree
MASTER OF SCIENCE
of ELECTRICAL ENGINEERING

By
Shoumik Roychoudhury
January, 2012

Thesis Approval(s):

Dr. Seong G. Kong, Thesis Advisor, Department of Electrical and Computer
Engineering, Temple University

Dr. Saroj Biswas, Committee Member, Department of Electrical and Computer
Engineering, Temple University

Dr. Joseph Picone, Committee Member, Department of Electrical and Computer
Engineering, Temple University

ABSTRACT

This thesis presents a multi-feature histogram approach to track a person in thermal vision. Illumination variation is a primary constraint in the performance of object tracking in visible spectrum. Thermal infrared (IR) sensor, which measures the heat energy emitted from an object, is less sensitive to illumination variations. Therefore, thermal vision has immense advantage in object tracking in varying illumination conditions. Kernel based approaches such as mean shift tracking algorithm which uses a single feature histogram for object representation, has gained popularity in the field of computer vision due its efficiency and robustness to track non-rigid object in significant complex background. However, due to low resolution of IR images the gray level intensity information is not sufficient enough to give a strong cue for object representation using histogram. Multi-feature histogram, which is the combination of the gray level intensity information and edge information, generates an object representation which is more robust in thermal vision. The objective of this research is to develop a robust human tracking system which can autonomously detect, identify and track a person in a complex thermal IR scene. In this thesis the tracking procedure has been adapted from the well-known and efficient mean shift tracking algorithm and has been modified to enable fusion of multiple features to increase the robustness of the tracking procedure in thermal vision. In order to identify the object of interest before tracking, rapid human detection in thermal IR scene is achieved using Adaboost classification algorithm. Furthermore, a computationally efficient body pose recognition method is developed which uses Hu-invariant moments for matching object shapes. An experimental setup consisting of a Forward Looking Infrared (FLIR) camera, mounted on a Pioneer P3-DX mobile robot platform was used to test the proposed human tracking system in both indoor

and uncontrolled outdoor environments. The performance evaluation of the proposed tracking system on the OTCBVS benchmark dataset shows improvement in tracking performance in comparison to the traditional mean-shift tracking algorithm. Moreover, experimental results in different indoor and outdoor tracking scenarios involving different appearances of people show tracking is robust under cluttered background, varying illumination and partial occlusion of target object.

ACKNOWLEDGMENTS

I would like to express my gratitude to my advisor Dr. Seong G. Kong for the support and direction that he has given me. He has spent many hours carefully revising and evaluating my research. I wish to thank my committee members Dr. Saroj Biswas and Dr. Joseph Picone for their helpful insights and suggestions. I would like to thank Dr. Youngjun Han as well as all the members of the Imaging and Pattern Recognition Lab for their assistance in completion of this thesis. I also extend my sincere thanks to faculty and students from College of Engineering in Temple University for their help. Finally, I wish to thank my family for their support and encouragement.

TABLE OF CONTENTS

ABSTRACT	ii
ACKNOWLEDGMENTS	iv
LIST OF TABLES	vii
LIST OF FIGURES	viii
CHAPTER	
1 INTRODUCTION	1
1.1 Object Tracking	1
1.2 Research Objective	5
1.3 Contributions	6
1.4 Thesis Organization	7
2 STATE-OF-THE-ART IN HUMAN TRACKING SYSTEM .	9
2.1 Object Representation	9
2.2 Detection and Sensors	12
2.3 Object Tracking	15
2.4 Introduction to Mean Shift Algorithm	16
2.4.1 Mean Shift Tracking	18
2.4.2 Drawbacks in Current state-of-the-art	21
2.4.3 Extensions in Mean Shift Tracking Algorithm	22
2.5 Applications in Human Tracking Systems	25
3 TRACKING WITH MULTI-FEATURE HISTOGRAM	33

3.1	Human Detection in Thermal Scene	33
3.1.1	Haar-like Features	35
3.1.2	Classification Algorithm	37
3.2	Body Pose Recognition	39
3.3	Multi-feature Histogram	43
3.4	Human tracking via Multi-feature Histogram	48
4	PERFORMANCE EVALUATION	53
4.1	Experimental Setup	53
4.2	Evaluation of Tracking System	55
4.3	Quantitative Results	57
4.4	Experimental Results	60
5	CONCLUSION	66
5.1	Future Work	68
	LIST OF REFERENCES	70
	BIBLIOGRAPHY	75

LIST OF TABLES

Table		Page
1	Extensions in mean shift tracking algorithm	23
2	Distance results for pose recognition	42
3	Characteristics of video sequences used for evaluation	57
4	Mean Absolute error(MAE) on OTCBVS benchmark dataset . .	60

LIST OF FIGURES

Figure		Page
1	Block diagram of human tracking in thermal vision	4
2	Hierarchy in object tracking	16
3	Flowchart of Human Tracking System	34
4	Flowchart of human detection module in thermal vision	35
5	Subset of the Haar-like features in human detection	36
6	Examples of a upright and rotated feature	36
7	Cascade of boosted classifiers with N-stages	38
8	Human detection results	39
9	Symmetrical body poses used for target identification	41
10	Recognition of “T” pose in body pose recognition module	43
11	Schematic diagram Normal Vs Multi-feature histogram	44
12	Normal vs Multi-feature histogram	46
13	Different edge contributions to multi-feature histograms of objects	48
14	Multi-feature histograms in a tracking sequence	51
15	Multi-feature histograms in thermal tracking sequence	52
16	Mobile robot setup	54
17	Examples from OTCBVS dataset	58
18	Tracking error plots on OTCBVS datasets	59
19	Tracking results in indoor environment(camera view)	61
20	Tracking of a single person	62
21	Tracking results in outdoor environment(camera view)	63

Figure		Page
22	Robust tracking under different pose orientations	64
23	Tracking in outdoor conditions	64
24	Tracking in the presence of multiple moving objects	65

CHAPTER 1

INTRODUCTION

This thesis addresses the moving object tracking problem which is a critical area of study in the domain of computer vision. It emphasizes on the development of a robust vision guided autonomous human tracking system that can be used to detect and track a person in surveillance like scenarios in both indoor and outdoor environments. The scope of this work is confined in the development of a software system that can be applied on commercially available open source research mobile robot platforms for development of state of the art surveillance systems. This thesis is part of an extended research where the goal is to infer information regarding human behavior and recognize human activity for threat analysis in a scene.

1.1 Object Tracking

Automatic moving object detection and tracking is a critical step in computer vision systems such as surveillance, perceptual user interfaces, augmented reality, smart classrooms, object based video compression and driving assistance. A robust tracking technique is essential in the context of successful retrieval of higher level object information. Information regarding the moving object of interest in terms of its location and shape of the object is necessary for object behavior recognition and analysis. Real time object tracking is a challenging problem as difficulties can arise from a variety of factors e.g abrupt change in object motion, changing appearance patterns of both the object and the scene, non-rigid object structures, object-to-object and object-to-scene occlusions and camera motion.

Object tracking can be defined as a way of estimating the trajectory of an object in an image plane as it moves around in the scene. A robust tracker assigns

a consistent label to the tracked object in every frame of the image sequences and depending on the problem domain, the tracker can provide object centric information such as orientation, size or shape of the moving object. Complexities in tracking processes arise due to several factors. There is always loss of information when the 3D environment is projected onto the 2D image plane. Information is also affected due to noise during image acquisition by the image sensors. Non rigid object shapes and complex object motions along with variation in object sizes and partial or full occlusion increases the complexity of the problem. Environmental factors such as variation in illumination condition also affect the tracking process. Moreover real time processing requirements is a big constraint in object tracking systems.

Due to the complex nature of the problems regarding the object tracking a lot of scientific research has been published for the past 20 years proposing numerous approaches. Each approach is specific for a particular application domain and no method can be generalized to give satisfactory results for all type of moving object tracking problems. The approaches differ on the basis of object representation, object features, modeling techniques of object motion, appearance and shape. Various methods have been proposed and improved starting from a simple and rigid object tracking under the conditions of a static camera to the more complex and non-rigid object tracking using non stationary camera. Yilmaz et al.[1] has a comprehensive study about different approaches developed in object tracking.

Object tracking in visible spectrum is very popular and it is quite natural to have numerous methods pertaining to object tracking in visible spectrum. However, all methods in visible spectrum have a common constraint of varying illumination conditions. Illumination variations quite understandably affect the performance of the object tracking algorithm due to change of appearance of the

object. Since camera sensors pertaining to visible spectrum depends on the optical property of light to generate the image of a scene, variations in scene illumination conditions change the appearance of both object and scene. Moreover in conditions such as very low illumination or complete darkness it is sometimes impossible to generate images using visible spectrum cameras. Thus it is sometimes difficult to use visible spectrum cameras in surveillance systems as constant monitoring is difficult in case of low illumination due to darkness or inclement weather.

Thermal infrared cameras helps to overcome the illumination problem of visible spectrum as thermal camera captures the thermal energy radiated by a moving object or person. Thermal cameras do not depend upon lighting condition for capturing of images so are useful in detecting objects in zero visibility conditions such as inclement weather and complete darkness. Thermal vision is also useful in detecting moving objects from cluttered background as cold non-living objects will have no thermal profile in the thermal IR spectrum thus warm moving objects can be easily be detected in the thermal IR spectrum. The Forward Looking Infrared (FLIR) camera has been used by the military to identify enemy forces. However with the increasing popularity of thermal vision and decreasing cost of FLIR camera systems, there has been a increase in interests among researchers about the possibility of using FLIR systems in civilian surveillance practices.

In this research the possibility of developing a *mobile surveillance* system has been investigated. By the phrase *mobile surveillance* we mean an autonomous vision guided robotic system that would be able to detect and track human beings in surveillance like conditions. As Research robots are gradually moving towards real world applications in populated environments, robust tracking algorithms are required to track people as well as other non rigid moving objects. Mobile surveillance is a useful way of threat detection and identification specially in

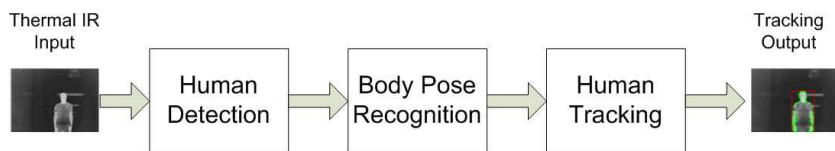


Figure 1: Block diagram of human tracking in thermal vision

places where human fail to identify threat or human involvement is impossible or dangerous. Traditionally tracking has been investigated for surveillance applications with the use of static cameras where simple background subtraction techniques can be employed to detect moving objects in the scene. In the case of mobile robots, the task of detection and tracking becomes more challenging since the robot is moving and as a result the background is dynamically changing. Thus traditional background subtraction techniques cannot be applied for moving object detection and tracking.

Most of the mobile robot systems used for detecting and tracking moving objects such as people use either range sensors such as laser scanners or a visible camera as the primary sensor. Both color vision and range sensor information have been combined to build tracking systems for mobile robot platforms. However performance of a tracker using features from the visible spectrum is highly dependent on illumination conditions of the surrounding region. As a result object regions with changing color information due to illumination variations or similar color profile with the background cannot be effectively tracked. In this thesis, thermal infrared camera (FLIR) is used as the only visual sensor to acquire information from the scene under investigation.

1.2 Research Objective

In this research a kernel based tracking technique known as Mean-shift algorithm has been adapted for tracking deformable moving object from a mobile robot platform using thermal vision. The goal is to track a single moving object in a cluttered background and extract higher-level information regarding tracked object for further object analysis and identification. The system detects moving human objects in the scene and tries to recognize human gestures in order to identify threat. Once the human gesture has been recognized the system locks on to that particular human object and starts tracking. A Forward Looking infrared camera is used which is mounted on a Pioneer P3-DX Mobile Robot platform. The vision is to develop a mobile surveillance system which can autonomously track moving object and identify threat in both indoor and uncontrolled outdoor environments. The image features in the visible spectrum are sensitive to illumination variation thus affecting the performance of the object tracking module. Also the system will need to track moving objects in uncontrolled environments where artificial illumination is not possible so features from visible spectrum would not prove to be a robust cue for tracking. On the other hand thermal infrared imagery which captures energy radiated from an object is robust to illumination variations so is a good alternative to visible cameras in the field of surveillance and object tracking specially in dark conditions where visible camera would fail to capture any image of the scene. A top level block diagram of the overall human tracking system is shown in figure1 This research can be broadly classified into three main modules.

Human Detection

For successful tracking of moving human objects in the scene from the mobile robot platform in thermal vision, robust detection of human is required for identification of humans. Human moving objects needs to be identified in the scene

and categorized as humans so that other moving objects are not misclassified as moving human objects. A classification algorithm based on Haar like features and Adaboost algorithm is used in this step to differentiate human from non-human objects.

Body Pose Recognition

In order to recognize human behavior for threat analysis and human activity recognition a human body pose recognition module is required where a known body gesture is recognized. The Hu-invariant Moment of the object contour is computed to estimate the body pose of the detected human object.

Human Tracking

The tracking algorithm is based on Mean-shift algorithm where the histogram of the target object is compared with the histogram of the current candidate object in each frame by maximizing a similarity function. Traditional mean-shift approaches used in visible camera use only color information to compute the histogram of the object of interest. In thermal infrared images due to low resolution images often only a single feature information is not sufficient for tracking objects as similar gray intensity level regions might be present in the background and due to lack of information the tracker will tend to drift away from the candidate object while tracking. To overcome this problem, a fusion of multi-feature information is used in this research to compute the histogram of the object of interest. Here along with the intensity level brightness information regarding the object's thermal profile, the edge information is also computed and both of these information are fused together to compute a multi-feature histogram of the combined information.

1.3 Contributions

Multi-feature histogram for increased robustness in mean shift object tracking in thermal vision

In this research the mean-shift object tracking procedure has been adapted to track the object of interest in the thermal scene. Traditional mean-shift tracking applications used a single feature information such as color information of the object to compute the histogram of the object. However in thermal imagery due to the low resolution of the thermal IR frames using only gray level brightness intensity information as a feature, tracking results are not robust in complex thermal scenes as the tracker is easily drifted away from the target when a object of similar brightness intensity is present in the scene which is quite common in outdoor tracking scenarios. To overcome this problem edge information is fused with gray level intensity information to create a multi-feature histogram using multi-feature information set. This proposed multi-feature histogram is constructed using this multi-feature information which is comprised of both gray level brightness intensity as well as edge information of the region of interest.

1.4 Thesis Organization

The rest of the thesis is organized as follows: Chapter 2 presents the state-of-the-art in people tracking including models and sensors used for detecting people, the theory behind kernel based mean-shift algorithm and its application to object tracking. This chapter also explains the the mathematical foundation of mean-shift algorithm. In the later part of the chapter the drawbacks of the current state-of-the-art in mean-shift algorithm is discussed and different approach to compute the histogram of the object region is suggested. The existing applications of people tracking is also discussed with a special focus on mobile robotics. In chapter 3 the multi-feature histogram based human tracking system developed in this research is introduced. The object identification procedure is discussed which focusses on human detection in a thermal scene and the object of interest selection

via a body pose recognition module. The details of multi-feature histogram is discussed. Furthermore, the tracking procedure using the multi-feature histogram of target object is presented. Chapter 4 introduces the experimental set-up, including the mobile robot and the thermal IR sensor, on which the entire system is implemented. The process of collecting ground truth data together with the performance metric used for performance evaluation of the tracking system is discussed. The experimental results are shown and the performance of the tracking system is evaluated. Chapter 5 concludes the current research. Limitations are discussed and further improvement are proposed for future work.

CHAPTER 2

STATE-OF-THE-ART IN HUMAN TRACKING SYSTEM

This chapter presents the state-of-the-art in people tracking and the theoretical basis for the people tracking system presented in this thesis. An overview of the most popular models and sensors used for detecting people is discussed in the beginning. Later the theory of object tracking covering mean-shift algorithm is presented. In addition the related work on people identification is briefly reviewed. Finally an overview of existing applications of people tracking with special focus on mobile robotics is presented.

2.1 Object Representation

Models of people help to solve two different kinds of problem in people recognition and tracking: to separate persons from other objects in the environment (detection) and to infer higher level understanding about the activity in the scene (recognition). The latter problem could be further decomposed, in increasing order of difficulty, into the problems of data association or deciding on a frame-by-frame basis (tracking) and absolute identification (identity recognition or activity recognition). This thesis is focused on the problem of detection and tracking; therefore only the problems of data association are considered. Further extensions allowing for absolute identification of persons, even though possible within the existing framework, would require incorporation of reliable recognition techniques, for example, face recognition or body pose estimation. The increasing complexity of these extensions would require more resources such as an increasing amount of memory (i.e., memory of recent frames, previous tracks and of all people in database) and computational power.

In detection the main difficulty is to extract common properties for all persons from the broad variety of human appearances. This appearance depends on a person's "size", "shape" etc. Moreover the appearance is affected by projection of the scene onto the sensor space, resulting in self-occlusions and occlusions by other objects and persons in the environment. In addition different individuals behave in different ways (standing, walking, sitting, lying down, cycling etc.) and their bodies can assume different poses. This also affects the detection task. On the other hand all these variations in appearance and behavior make the identification task possible. The main goal in this case is to find specific and invariant properties for each individual. Therefore the choice of a proper person model for a specific application will always be related to the type of application being developed.

Another important issue that should be discussed is the complexity of the model. Complex models can provide very detailed information that is required in applications such as simulating virtual agents, or systems analyzing the movement of a sportsman or dancer. Such systems usually do not have strong constraints about processing time, often working in an off-line manner, and allow for special arrangements of the environment. In contrast on-line systems such as mobile robots usually do not require such detailed information, and therefore tend to favor simpler models that can fulfill the strict requirements for processing speed and robustness. Therefore the complexity of a model will be dictated by the demands of a specific application limited by the available resources (sensors and computational power).

Let us present some of the existing models used in people recognition systems. We will use a general classification that separates them into object-centered and view-centered models. Object-centered (also called view-independent) models are based on the structural characteristics of a person that are invariant to different

view-points. Depending on the representation these models can be categorized into stick figures [2] and volumetric models [3]. Stick figures represent the skeletal structure of the body while volumetric models attempt to represent the whole body by decomposition into basic geometrical shapes such as spheres or cylinders.

Object-centered models are used mostly in recognition tasks that require more complex analysis of the human body (e.g. gait recognition). One serious drawback of these models is the fact that they require a pose recovery procedure that maps information provided by the sensors to a 3D representation. This task is often computationally complex and demands special conditions such as use of multiple sensors and/or markers mounted on the person's body. View-centered models (or appearance models) are grounded in features extracted from the information provided by sensors. These features correspond to different appearances of a person due to, e.g., different view-points, light conditions, poses of the body, etc. Existing approaches use features such as points, edges, blobs [2], [4]. View-centered models avoid the difficult pose recovery step required by object-centered models. This fact makes view-centered models more robust in real time systems.

Moreover appearance models are not restricted to 2D information but may also contain 3D information (obtained from e.g., stereo-vision, structure from motion, range sensors, etc.). From the perspective of mobile robotics, appearance models are more desirable since they are directly grounded in the robot's perception (there is no need to find correspondences between model components and image features). The internal representation in the sensor space does not limit possible applications and tasks (e.g. person following, user recognition). In general appearance models are also more robust and require less computational power, which in the case of limited hardware resources of a robot and high real-time demands cannot be ignored. In this thesis we use a simple appearance model that approximates a

person’s projection onto the image space. Its simplicity allows this model to be combined with a fast tracking method. The model is based on thermal information which allows robust tracking of persons even in darkness. Our model helps to solve the two problems of detection and tracking.

2.2 Detection and Sensors

Traditionally people detection is considered as a task carried out before tracking that determines the presence and number of persons from the input sensory data. This is realized by segmentation of the image data into regions corresponding to each detected person, usually by use of some model of a person. In this section we present the most popular sensors and methods used to detect people, with a special focus on mobile robotic applications.

The most popular sensors used for detecting people are visible cameras. Most existing vision-based methods concern non-mobile applications (e.g. surveillance, pedestrian detection) where the pose of the camera is fixed. Detection in this case can be solved by background subtraction Haritaoglu et al. [5] or temporal difference [6]. In the first method foreground objects in the image frame are segmented after subtraction of the background model of the scene. The temporal difference method uses differences between two consecutive frames to determine moving objects. Both approaches make a strong assumption that detected objects are persons. Other techniques use a further recognition step in which persons are discriminated from other objects [7] [8].

Techniques based on skin color can be used regardless of the motion of the sensor, therefore being very popular in mobile robotics applications [9]. The skin color of the human body is quite unique compared to other objects, which allows segmentation of regions in the image corresponding to the face or hands of a person. Similar approaches for detecting humans are based on face detection

algorithms. Some popular methods from the vast variety of different algorithms include principal component analysis (PCA) [10], template matching, or rapid detectors by Viola and Jones [11]. However, methods based on skin color or face detection are usually limited to face and hand detection, hence persons must be facing the sensor. Recent advances in visual object recognition provide learning techniques that enable detection of people without assuming any a priori knowledge of the scene. They are, however, computationally demanding. All of the above mentioned vision-based systems share common problems such as shadows, varying lighting conditions and occlusions.

Use of non-standard vision sensors for people detection such as a stereo camera [12] or thermal sensor [13] helps to overcome some of the problems related to color vision. Stereo vision provides extra range information that makes segmentation easier, allowing for detection of both standing and moving people. Stereo vision has been applied only in a few mobile robotic applications [14] [15], perhaps due to the low resolution of depth information available from these sensors (typical stereo vision systems quantize the depth estimates into a maximum of 32 layers/disparities). Thermal vision takes advantage of the fact that humans have a distinctive thermal profile compared to nonliving objects. Moreover thermal information is not influenced by changing lighting conditions and allows detection of people even in darkness. Infrared sensors have been applied to detect pedestrians in a driving assistance system: Bertozzi et al.[16] use a template based approach while Nanda and Davis [13] apply different image filtering techniques. Meis et al. [17] filter the whole image and classify persons based on the symmetry of detected gradients. Xu et al. [18] employ a classification method based on a support vector machine. However till now there is hardly any published work on using thermal sensor information to detect humans on mobile robots. The main reason for the

limited number of applications using thermal vision so far is probably the relatively high price of this sensor, which is gradually decreasing.

Other types of sensors that can be used for people detection include range-finder sensors such as laser and sonar. These are very popular sensors in mobile robotics for navigation and localization tasks [19]. A system described by Schulz et al. [20] detects local minima in range readings caused by the legs of a person and then removes all static objects by subtracting consecutive laser readings. In the paper by Kluge et al. [21] the authors cluster scan data into a set of points representing objects and by performing shape analysis extract those points corresponding to people. Both approaches detect moving objects rather than people. Despite the limitations of systems based on laser scanners (i.e. they can only detect moving objects rather than humans), they remain popular sensors in mobile robotic applications because of the low computational demands due to the low dimensionality of sensor data. Recent progress in building 3D range sensors makes them promising sensors for future applications requiring people detection. To overcome some problems related to a specific sensor it is possible to combine information from different sensors. For example, Feyrer and Zell [22] use different features provided by a color and stereo camera together with a laser scanner, and Wilhelm et al. combine color vision with sonar. This approach generally leads to more robust recognition systems. However, another problem arises here, namely sensor fusion how to combine the different types of sensor information.

Our mobile robotic system uses a thermal camera to efficiently detect persons despite the motion of the platform. The distinct thermal profile of the human body is used to along with the edge information to create a multi-feature histogram of the person detected via boosted classification technique for people detection in thermal vision.

2.3 Object Tracking

Object tracking is one of the most fundamental problems in the computer vision domain and it has always been and still is an important area of research. Tracking algorithms are normally evaluated by their abilities to handle the different complexities that are accompanied by an object tracking problem as mentioned in chapter 1. Yilmaz et al. [1] categorize the tracking methods into 3 groups named **point tracking**, **kernel tracking** and **silhouette tracking**.

For point tracking methods, an object is represented by a number of points, and the correspondence of these points are tracked over consecutive frames. The points are normally combined together with a model, and the correspondences can be evaluated over a number of constraints, such as spatial constrain or motion model. Particle filter, Kalman filters and its extension like Extended Kalman filters are some of the most common point tracking algorithms. In kernel based tracking methods, an object is normally represented by primitive geometry shape, and the motion of the object region can be estimated over consecutive frames. The kernel based tracking algorithm can be much different due to the usage of different object representation methods and motion estimation methods. When the object has more complex shape, a primitive region is not enough to model the object. In this case silhouette tracking methods are normally applied. Common silhouette tracking algorithms include shape matching and contour tracking. In case of contour based methods like snake[23] or level set[24] are mainly used to track object contours. Yilmiaz [25] incorporated prior shape into the objects energy and used level set to evolve the contour by minimizing the energy functional. In order to track objects with non-Gaussian state density in cluttered background, Isard and Blake [26] presented condensation algorithm. Although contour based tracking

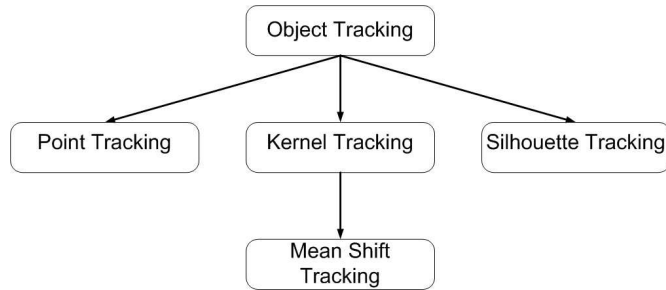


Figure 2: Hierarchy in object tracking

can achieve high precision tracking, however their robustness is usually not better than region based kernel tracking methods. Also the computing cost of object contour is very high, specially for fast moving objects. In this thesis, the proposed human tracking system is based on mean shift tracking algorithm which belongs to the group of kernel tracking algorithms.

2.4 Introduction to Mean Shift Algorithm

The mean shift procedure is a non-parametric statistical procedure for seeking the nearest mode in a sample distribution. It was originally presented in 1975 by Fukunaga and Hostetler [27] and later by Cheng [28] in 1995. In this procedure the maxima of a density function is located from the given discrete data, sampled from that function. It is used to identify the modes in the density function. A good introduction of mean shift procedure can be found in [29]. If there is a set $S = \{\mathbf{x}_i\}_{i=1,2,\dots,n}$ of n points in d -dimensional space (R^d), the probability density function or multivariate kernel density estimate using kernel $K(\mathbf{x})$ and window size h at point \mathbf{x} is given by

$$\hat{p}_K(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^n K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) \quad (1)$$

$K(\mathbf{x})$ is a multivariate kernel function with a window radius of h . It has been proven that if $K(\mathbf{x})$ is isotropic kernel with a convex and monotonic decreasing

profile then the mean-shift vector always points in the direction of the maximum increase in the density and thus following the direction recursively leads to the mode of the density function spanned by S . The gaussian and the Epanechnikov kernels are examples of such kernels. In [30] we find the related proof of convergence. Equation 1 can be inferred from the Parzen window technique [31] of probability density estimation. $K(\mathbf{x})$ is radially symmetric, and the profile function of kernel $K(\mathbf{x})$ for $\mathbf{x} > 0$ is given by

$$K(\mathbf{x}) = k(\|\mathbf{x}\|^2) \quad (2)$$

Applying equation 2 to equation 1 we have:

$$\hat{p}_K(\mathbf{x}) = \frac{1}{nh^d} \sum_{i=1}^n k\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right) \quad (3)$$

We also define another kernel $G(\mathbf{x})$ with profile $g(\mathbf{x})$ where,

$$g(x) = -k'(x) \quad (4)$$

This kernel $G(\mathbf{x})$ is called the shadow of the kernel $K(\mathbf{x})$.

Now the estimate of density gradient can be obtained as the gradient of the estimated density:

$$\hat{\nabla} p_K(\mathbf{x}) \equiv \nabla \hat{p}_K(\mathbf{x}) = \frac{2}{nh^{d+2}} \sum_{i=1}^n (\mathbf{x} - \mathbf{x}_i) k'\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right) \quad (5)$$

recalling the equation 4 the above equation can be rewritten as:

$$\begin{aligned} \nabla \hat{p}_K(\mathbf{x}) &= \frac{2}{nh^{d+2}} \sum_{i=1}^n (\mathbf{x}_i - \mathbf{x}) g\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right) \\ &= \frac{2}{nh^{d+2}} \sum_{i=1}^n g\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right) \left[\frac{\sum_{i=1}^n \mathbf{x}_i g\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^n g\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right)} - \mathbf{x} \right] \\ &= \frac{2}{h^2} \hat{p}_G(\mathbf{x}) \left[\frac{\sum_{i=1}^n \mathbf{x}_i g\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right)}{\sum_{i=1}^n g\left(\left\|\frac{\mathbf{x} - \mathbf{x}_i}{h}\right\|^2\right)} - \mathbf{x} \right] \end{aligned}$$

Thus the *mean shift vector* $m(\mathbf{x})$ can be defined as:

$$m(\mathbf{x}) = \frac{\sum_{i=1}^n \mathbf{x}_i g(\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\|^2)}{\sum_{i=1}^n g(\|\frac{\mathbf{x}-\mathbf{x}_i}{h}\|^2)} - \mathbf{x} \quad (6)$$

Therefore it can be rewritten $m(\mathbf{x})$ as:

$$m(\mathbf{x}) = \frac{h^2 \nabla \hat{p}_K(\mathbf{x})}{2\hat{p}_G(\mathbf{x})} \quad (7)$$

Thus it can be concluded that at location \mathbf{x} , the mean shift vector estimated with kernel $G(\mathbf{x})$ is proportional to the density gradient estimated with kernel $K(\mathbf{x})$ normalized by the density estimated with kernel $G(\mathbf{x})$. In other words the mean shift vector points to the direction where the density will be maximum.

In more generic terms the mean shift procedure can be defined in two steps:

1. Calculate the *mean shift vector* using kernel G according to equation 7
2. Move \mathbf{x} according to the *mean shift vector*

These two steps are repeated until the *mean shift vector* vanishes or \mathbf{x} reaches the mode of the distribution where the estimated density with kernel K is maximized.

2.4.1 Mean Shift Tracking

In [30], the author took the advantage of the mean-shift's property of locating local maximum in a iterative gradient ascent procedure and proposed an elegant method to track blobs based on intensity histogram. The essence of this tracker is region matching. The match criterion is a similarity measure based on colored histogram. Between frames, the region being tracked can change location, appearance and size. To examine possibilities quickly, the mean shift algorithm is used as a way to converge from an initial guess for location and scale to the best match based on the color-histogram similarity.

The histogram is computed over an ellipsoid window, and the contributions

at each point are gaussian weighted based on their distance from the window's center. An initial, hand-positioned ellipse defines the target distribution. Location is updated in each frame using the mean-shift algorithm across pixel location and scale to find the local maximum for similarity to the target distribution. Similarity measure is done using the Bhattacharyya coefficient [30] or Kullback-Leibler divergence.

The mean shift based tracking algorithm is a 2 steps bottom-up approach: first, object is represented; second, object is located in consecutive image frames. In the mean shift tracking algorithm, an object is represented by weighted histogram within the object region. The object region or shape can be an ellipse or a rectangle. The weight is assigned in a way that the pixels nearer to the center of the object get bigger weight, because they are normally more reliable than the outer ones.

If $\hat{\mathbf{q}}$ is defined as the object reference then it can be said that

$$\hat{\mathbf{q}} = \{\hat{q}_u\}_{u=1\dots m} \quad (8)$$

where m is the number of bins in the histogram and u is the index. If we consider the center of the object is (x, y) , and the size of the object is $h_w \times h_h$ rectangle, then we can have:

$$\hat{q}_u = C \sum_{i=1}^n k\left(\left\|\frac{x_i - x}{h_w}, \frac{y_i - y}{h_h}\right\|^2\right) \delta[b(x_i, y_i) - u] \quad (9)$$

where C is a normalization constant to fulfill the constraint:

$$\sum_{u=1}^m \hat{q}_u = 1 \quad (10)$$

and $b(x_i, y_i)$ is function to return the histogram index of the feature value at (x_i, y_i) .

Similarly, an object candidate with size $h'_w \times h'_h$ centered at position (x', y') can be defined as $\hat{\mathbf{p}}(x', y')$:

$$\hat{\mathbf{p}}(x', y') = \{\hat{p}_u(x', y')\}_{u=1\dots m} \quad (11)$$

$$\hat{p}_u(x', y') = C' \sum_{i=1}^{n_h} k\left(\left\|\frac{x_i - x'}{h'_w}, \frac{y_i - y'}{h'_h}\right\|^2\right) \delta[b(x_i, y_i) - u] \quad (12)$$

where C' is a normalization constant to fulfill the following constraints:

$$\sum_{u=1}^m \hat{p}_u(x', y') = 1 \quad (13)$$

Object localization is actually to find a location where the similarity of the object candidate and the object reference is maximized. Bhattacharyya coefficient is used to measure the likelihood in mean shift tracking algorithm

$$\rho[\hat{\mathbf{p}}(x, y), \hat{\mathbf{q}}] = \sum_{u=1}^m \sqrt{\hat{p}_u(x, y) \hat{q}_u} \quad (14)$$

If the object candidate $\hat{p}(x, y)$ does not change drastically from the initial $\hat{p}_u(\hat{x}_0, \hat{y}_0)$, $\hat{\rho}(x, y)$ approximates:

$$\rho[\hat{\mathbf{p}}(x, y), \hat{\mathbf{q}}] \approx \frac{1}{2} \sum_{u=1}^m \sqrt{\hat{p}_u(\hat{x}_0, \hat{y}_0) \hat{q}_u} + \frac{1}{2} \sum_{u=1}^m \hat{p}_u(x, y) \sqrt{\frac{\hat{q}_u}{\hat{p}_u(x_0, y_0)}} \quad (15)$$

Applying definition of the \hat{p}_u to the above equation, it becomes:

$$\rho[\hat{\mathbf{p}}(x, y), \hat{\mathbf{q}}] \approx \frac{1}{2} \sum_{u=1}^m \sqrt{\hat{p}_u(\hat{x}_0, \hat{y}_0) \hat{q}_u} + \frac{C_h}{2} \sum_{i=1}^{n_h} \left[\sum_{u=1}^m \sqrt{\frac{\hat{q}_u}{\hat{p}_u(\hat{x}_0, \hat{y}_0)}} \delta[b(x_i, y_i) - u] k\left(\left\|\frac{x_i - x}{h'_w}, \frac{y_i - y}{h'_h}\right\|^2\right) \right] \quad (16)$$

If we assign

$$w_i = \sum_{u=1}^m \sqrt{\frac{\hat{q}_u}{\hat{p}_u(\hat{x}_0, \hat{y}_0)}} \delta[b(x_i, y_i) - u] \quad (17)$$

Finally it can be written as

$$\rho[\hat{\mathbf{p}}(x, y), \hat{\mathbf{q}}] \approx \frac{1}{2} \sum_{u=1}^m \sqrt{\hat{p}_u(\hat{x}_0, \hat{y}_0) \hat{q}_u} + \frac{C_h}{2} \sum_{i=1}^{n_h} w_i k\left(\left\|\frac{x_i - x}{h'_w}, \frac{y_i - y}{h'_h}\right\|^2\right) \quad (18)$$

It is to be noted that the first term of the equation 18 is constant once initial pixels are chosen, and the second term is actually the density estimate computed with kernel $K(x, y)$, with the data being weighted by w_i . Then recalling the mean shift theory, the likelihood $\rho[\hat{\mathbf{p}}(x, y), \hat{\mathbf{q}}]$ can be maximized by iteratively moving (\hat{x}_t, \hat{y}_t) to the new position $(\hat{x}_{t+1}, \hat{y}_{t+1})$.

2.4.2 Drawbacks in Current state-of-the-art

Mean-shift tracking is one of the popular methods of object tracking and no doubt has advantageous over more common approaches such as template matching. Template matching is a brute force method of searching the image for a region similar to the object template defined in the previous frame. This brute force method has high computational cost. In mean-shift tracking method the histogram of the object of interest is computed and used as a target model for object tracking. A weighted histogram is computed from a window region and instead of applying brute force search for locating the object a iterative gradient ascent procedure is adapted where a similarity function is maximized by comparing the histogram of the target model with the histogram of the hypothesized target candidate. Then mean shift tracking algorithm maximizes the appearance similarity of the object target and object candidate iteratively by comparing the weighted histogram of the object \hat{q} and the window around the hypothesized object location, \hat{p} . The similarity metric is defined based on the Bhattacharya coefficient. At each iteration, the hypothesized object location is moved according to the mean shift vector, where the similarity of the hypothesize object and the target object is increased. The biggest advantage of mean shift tracking is that the computational cost is much cheaper than other matching method because the dense gradient climbing approach is used rather than the brute fore searching approach. Therefore, it became one of the most popular object tracking algorithms for its low computational cost and robustness.

However there are few limitations to the original implementation for object tracking. Firstly, the spatial information of the object is not strongly encoded in the representation of the object, thus the scale and orientation information of the object will be lost during tracking. Comaniciu et al. [32] gave an solution that

one can search the nearby scale to find most fit scale, however it is not sufficient in a complex situation when the object changes its scale rapidly. Due to the usage of circular kernel, the mean shift tracking algorithm is also invariant to object orientation change. Secondly, the mean shift tracking algorithm uses a static model of the object which assumes that the object will not change its outlook much which is not true in the real environment. For example, one can easily fail a mean shift tracker by rotating the tracked object to the other side (suppose that the two sizes of the object are different which is true for most of the cases). Moreover, there are also other constraints for mean shift tracking like it assumes that the object will not move more than its own size between 2 consecutive frames, thus searching window size is limited to the size of the object. This decreases the computational cost and the distraction of the background, but makes it less robust for the case of fast object motion. Increasing the searching window size results in robustness for fast object motion but the computational cost and the background distraction problem will be faced.

2.4.3 Extensions in Mean Shift Tracking Algorithm

To tackle the mentioned problems, researchers have proposed many extensions of the original mean shift tracking algorithm. They can be broadly divided into 3 categories: Use of a new object representation, use of new similarity measure functions and updating the target model. A category of these algorithms are shown in table 1

Collins [33] adapted Lindeberg's theory [41] of feature scale selection based on local maximum of differential scale-space filters to the problem of selecting kernel scale

Table 1: Extensions in mean shift tracking algorithm

Algorithms	object representation	Scale invariance	Rotation invariance	Similarity measure function	Target model Update
Collins[33]	✓	✓			
Zhang[34]	✓	✓	✓		
Jamil and Sebastien[35]	✓		✓		
Leung and Gong[36]	✓				
Bradski[37]	✓	✓	✓	✓	
Yang et al.[38]				✓	
Comaniciu and Ramesh [39]					✓
Peng et al.[40]					✓
The proposed algorithm	✓	✓			✓

for mean-shift blob tracking. The features are extended from 2D spatial space into a 3D spatial-scale space by applying a different of Gaussian (DOG) kernel (with different covariance) on the image, and then the mean shift tracking is carried out in the spatial-scale space, where both the optimal location and scale of the object can be achieved simultaneously. This method gave good result for tracking objects with scale change, however it is computational expensive, because of the construction of the scale space and the increased dimensionality for mean shift. Moreover, it is still invariant to rotation. Based on the work of Collins, Zhang et al. [34] proposed a new approach where both the scale and orientation of the object can be tracked. Similar to Collins method, a 2D ellipse regularization log kernel is used for scale selection while another 2D ellipse Gaussian kernel with orientation information is used to not only describe the foursquare characteristic but also provide the orientation information of the characteristic. In their method, the location, orientation and scale information are tracked sequentially. Rather than constructing a scale or rotation space, Jamil and Sebastien [35] used a simple but efficient method to track object orientation. For every object they build a gradient histogram to encode the orientation information. In the beginning of the

tracking, the image is rotated into different angles in order to have the gradient histogram of the object with different orientations and during tracking the most matching histogram indicates the orientation of the object. The computational cost of the gradient histogram is not expensive, and effectiveness of the algorithm can be achieved. However, only limited number of orientations is available and as the angles is increased the computational cost also increases. In another paper, Leung and Gong [36] introduce random sampling into mean shift tracking algorithm. The object is represented by selected samples rather than the whole image region, where the computational cost is reduced significantly without losing much robustness. However an online feature selection mechanism is required. This method can be extremely useful when working with high resolution image and big size object tracking. In the paper of Bradski [37], a similar mean shift tracking method called CAMSHIFT was described even earlier than Comaniciu. Comparing to mean shift tracking algorithm, CAMSHIFT uses histogram back-projection as the measurement of the similarity between the object target and candidate. Moreover, the distribution of features was used to estimate the scale and orientation. However, the algorithm was designed to the problem of face tracking where many assumptions like skin color face shape were made. Therefore the performance on normal object tracking in a complex environment may not be sufficient. In 2005 Yang et al [38] exploit a new sample based similarity measure for mean shift tracking algorithm instead of the Bhattacharyya coefficient based similarity metric used by Comaniciu. Instead of evaluating the information-theoretic measures from the estimated pdf, they directly define the similarity between two distributions as the expectation of the density estimates over the model or target image. The new measurement achieves higher discriminative and provide a better outlier rejection property. In paper [39] Comaniciu and Ramesh

showed how the mean shift tracking can be combined together with Kalman filter where the Kalman filter was used to predict the next state of the tracked object. Later Peng et al. [40] demonstrated another usage of Kalman filter, which is to update the object model adaptively during the tracking. This makes the tracking more robust to the object appearance change.

Furthermore all original implementations of mean-Shift algorithm use a single feature for histogram computation which is mainly using the color information of the object. The grey level intensity information of the object has been a the primary feature for target tracking in infrared images. However, this feature is not a robust source of information when it comes to target tracking in thermal vision because it is vulnerable to disturbances of similar grey background regions in the scene. In this thesis a more robust multi-feature set has been utilized for the computation of histogram in the mean-shift tracking framework. The object edge information is fused with the grey level intensity information of the object to model that the human body structure. This multi-feature target model provides robust tracking results in IR sequences.

2.5 Applications in Human Tracking Systems

Many human tracking systems have been designed for applications such as surveillance, video games, virtual reality interfaces, gesture recognition, user interfaces etc. The majority of these systems use a stationary camera and are usually designed to work in known or partially controlled environments, which allows simplification of the detection task. The most popular models used are appearance models of humans, however in applications for movie industry or motion analysis more complex 3D models are used. Real-time systems usually use tracking techniques aided with heuristics to simplify the problem. Systems working in an off-line manner apply state-of-the-art tracking algorithms allowing

to handle multiple persons and occlusions.

PFinder is one of the first people tracking systems developed [4]. The system was used in many successful applications such as video games, virtual reality interface or even for gesture recognition. The system is able to track a single person in real-time (10 fps, 160x120 pixels). PFinder uses an appearance model based on statistics of color and shape. This model is learned before-hand from data. Different body parts of a person are represented as blobs. This representation allows also for recognition of simple gestures. PFinder uses a stationary camera and background subtraction to detect a person and initialize the respective model. Tracking is realized by predicting position of blobs based on a constant velocity motion model and later updating the respective models based on the classification of pixels by a maximum a posteriori approach. The system W4 proposed by Haritaoglu et al. [5] is similar in spirit with PFinder. It also uses a blob based representation of the human body and learns appearance models based on histograms. The system can also estimate the rough pose of the body. The most important difference to PFinder is that W4 system can track multiple persons: single separated persons and groups. The system uses a heuristic approach to tracking based on combination of prediction from the second order motion model together with correlation techniques updating the model. The Reading People Tracker also allows to track multiple persons. The system uses a Kalman filter based active shape tracker to track detected persons and has the ability to deal with partial occlusions. Smith et al. [42] presents the latest achievements in the field of people tracking based on stationary cameras. The system is using a particle filter based multi person tracker. It models interactions between occluding persons hence is able to deal with partial and total occlusions. The advanced tracking methods do not allow to implement the system in real time, however Gavrilu and

Davis [43] and Sidenbladh [44] are examples of systems that use complex 3D models of persons. Because of their computational complexity they can be used only in specific applications without real-time requirements such as in movie industry or motion analysis of the human body. Both systems allow to track a single person only. The first system in addition requires a specially controlled environment (i.e. two camera set-up, persons wearing tight clothes). The latter system uses a particle filter to robustly track the pose of a person.

In the case of mobile robots, human tracking becomes an even more challenging task because the robot is moving and the environment is unpredictable. In addition computational resources are very limited since a robot usually has to perform other tasks such as navigation, planning, object recognition, etc., at the same time. Therefore not all techniques used in non-mobile applications can be directly applied to mobile robotics. Some of the existing mobile robotic applications designed to detect and localize people in the environment is presented in this section. Here the literature has been divided into systems which only detect humans based on the current sensor data and systems which track them using recursive methods for state estimation. Otherwise it is difficult to classify existing systems into distinct categories, because the field is at an early stage and a wide variety of techniques are still being explored. Perhaps two general trends can be observed: First, a few approaches rely on a model of the environment for separating humans from the background, therefore requiring accurate maps and self-localization by the robot. Second, many approaches apply complementary sensor modalities, e.g., vision and range-finder data, in order to compensate for the limitations of each individual modality. For example, many approaches try to detect human legs from local minima in laser scans, but usually this information alone is not enough to guarantee reliable results, and another sensor (e.g., vision)

may be used to confirm the presence of humans. In this thesis we use only thermal vision sensor for data capture and tracking purposes.

Early mobile robotic systems, despite their simplicity and hardware limitations, showed that people can be detected and localized in the environment even though the platform is moving. They usually made strong assumption about the environment and used very limited and simple models of persons, expecting the user to be somehow aware of the robot. The most popular sensor was a color camera working in low resolutions to simplify the image processing. They usually could detect only a single person since no tracking procedure was applied. Most probably the first robot designed to recognize people was Polly [45], a mobile robot that gave tours in the corridors of the MIT AI Lab. It was equipped with a simple vision system (a camera pointing down at the floor with a resolution of 64x48 pixels and frame rate of 15 Hz) that was capable of detecting people in its surroundings. The system could estimate the “depth” of different objects in the environment by filtering out pixels belonging to the floor. This approach assumes that the environment is planar, so that depth can be estimated from height in the image plane, and that the floor has a distinctive texture that can be easily separated from foreground objects. Based on this depth information Polly could detect objects corresponding to people’s legs (it was assumed that other similar objects like table or chair legs would not be present). Walls and junctions were detected by the same vision systems using a similar approach. Moreover the system could also recognize simple gestures such as foot waving, allowing simple interaction with the robot by the user. Blackburn and Nguyen [46] in 1994 presented a mobile robot equipped with a biologically inspired vision-control system that could separate the motion of a moving object from the motion of the environment caused by the movement of the robot. The reported speed of the system was 15 fps with images of resolution

128x128 pixels. The system required the speed of the tracked object to be high enough to separate the object from the background, so the approach would only be useful for tracking humans while they move quickly from one place to another.

Huber and Kortenkamp [14] presented a mobile robot with a visual attention system that was able to detect and follow an arbitrary object. The system used both stereo and motion information to detect the first object with enough texture information (assumed to be a person). This information was later used by the robot to pursue the detected object. The system could operate at a speed of 30 fps. Later in Kortenkamp et al. [47] a similar system was used for gesture recognition where a simple model of a person was introduced to obtain more reliable detection. Another robotic system recognizing gestures of a person was presented by Kahn et al. [15] in 1996. The robot Rhino and its successor Minerva are examples of very successful mobile platforms designed to work in museums as artificial tour guides.

Schlegel et al. [48] presented a people tracking system that uses a model of a person including a color histogram and adaptive contour model that is learned during an initialization phase. The system can detect and track people in a range of 1 to 5 m. In this system tracking is performed by an adaptive procedure that updates both models. Example of systems where tracking was realized in a similar way include Waldherr et al. [49] and [9]. Both systems use skin color to select the regions of interest and later update the color models of these regions. Wilhelm et al. present a robotic shop assistant. The tracking system is based on a particle filter allowing to track a single person (a potential user). The system uses a combination of skin color, contour based information and additional range information from the robot's sonar. Information from different modalities is combined by a fuzzy data fusion technique. Kleinhagenbrock et al. [50] 2002 present another technique that combines skin color and laser data. In this case the two modalities are fused by

means of symbols and a respective set of rules. The system is able to track one person and the reported speed is around 3-4 fps with resolution 198×139 pixels for the vision system tracking skin color and 4.6 Hz for the laser. The system is able to combine the different sensor data asynchronously.

The system described by Jensen and Siegwart in 2003 uses a laser scanner and is similar in principle to Montemerlo et al. system published in 2002 in the sense that the system uses a map of the environment and then selects objects that are outliers. This is realized by the EM algorithm and a feature-based representation of the environment to reduce the complexity of the algorithm. However, the EM algorithm requires many iterations and all previous data to be available, so the scalability of this approach is not clear. Another laser-based people tracking system mounted on a wheelchair is presented by Kluge et al. in 2001. Shapes of persons are represented by a set of objects (i.e. vertices and edges) extracted from the laser scan. Tracking of multiple moving persons is realized by matching objects from two consecutive scans represented in a graph-like structure. The association is performed by standard optimization techniques used in graph theory. However this approach allows to track only well separated, moving objects and data association realized by the graph can fail in cases of simultaneous appearance/disappearance of objects. Cielniak et al. in 2003 developed a method for people tracking by mobile robots allowing the incorporation of additional knowledge about the behavior of the people. These behaviors are learned off-line by clustering recorded trajectories of the persons provided by a laser scanner using the EM algorithm. The results from the clustering are later used to construct a person-specific hidden Markov model (HMM). This model can predict the intentions of a person and is used for on-line tracking. Range and color information is used to update the HMM model where color allows to distinguish between different individuals and tracking

of multiple persons. The issues related to on-line learning of behaviors would need to be investigated further to allow application of this system in practice. Zajdel et al. in 2005 addressed the problem of tracking and identification of persons from a mobile platform using vision. The proposed system segments persons from the image in two ways: by a standard background subtraction method when the robot is stationary and motion extraction from optical flow when the robot is moving. The tracking procedure is realized by a color matching algorithm. The resulting tracks are later used for re-identification of persons entering the field of view of the camera. This so-called global tracking is used to determine whether the observed person has been seen before. The method uses a Bayesian network that associates local tracks using color and spatio-temporal features extracted from the tracks. However the authors did not investigate or discuss some possible problems of the approach. This would include, for example, the influence of faulty tracks on the performance of the global matching algorithm, or tractability of the approach with a growing number of observed tracks (the complexity of the proposed Bayesian network grows exponentially with the number of tracks).

Human tracking systems have been used in many interesting applications in different fields. Not all of the techniques used in non-mobile applications can be directly transferred to robotic systems due to the increased amount of noise, movement of the platform, unpredictability of the environment and computational limitations. The presented mobile applications illustrate the need for fast and reliable human tracking systems.

In this chapter we presented the theoretical background and state of the art in human tracking for mobile robots. The major challenge for human tracking systems lies in reliable detection and localization of people. The selection of an appropriate model of a person depends heavily on the application, however in

general appearance models seem to be more suitable for mobile robots. The most popular sensors are vision cameras, but other sensors such as a thermal camera or laser scanner can simplify or aid the detection task.

CHAPTER 3

TRACKING WITH MULTI-FEATURE HISTOGRAM

This chapter presents a human tracking system developed that uses a multi-feature histogram in the mean-shift framework for successful tracking of the object of interest. Figure 3 shows the flow diagram of how the control of the program is transferred to the human tracking module once a certain human body pose is recognized. The human detection and the human body pose recognition module is introduced and discussed. It is necessary to identify the object of interest that needs to be tracked in the thermal scene. Successful identification procedure is necessary for tracking the object of interest in surveillance situation specially in threat monitoring scenarios. In this thesis the human object that needs to be tracked is identified via a body pose recognition. Furthermore, the multi-feature histogram is introduced and the details are presented regarding its usage in the mean shift framework to increase the robustness of the tracking capability.

3.1 Human Detection in Thermal Scene

The goal of object detection module is to find an object of a predefined class in static or video frames. The task is usually handled by extracting certain image features and then heuristics are applied to combine those feature characteristics to uniquely detect the object of interest. Detecting people is a complex task as it is hard to find unique characteristic features that can handle the huge variety of instances of human pedestrian object class. The size, color and style of clothing vary causing different appearances. Cluttered background in street scenes with the presence of other moving objects like cars poses threat of miss detection. Illumination and inclement weather conditions cause distortions in image features.

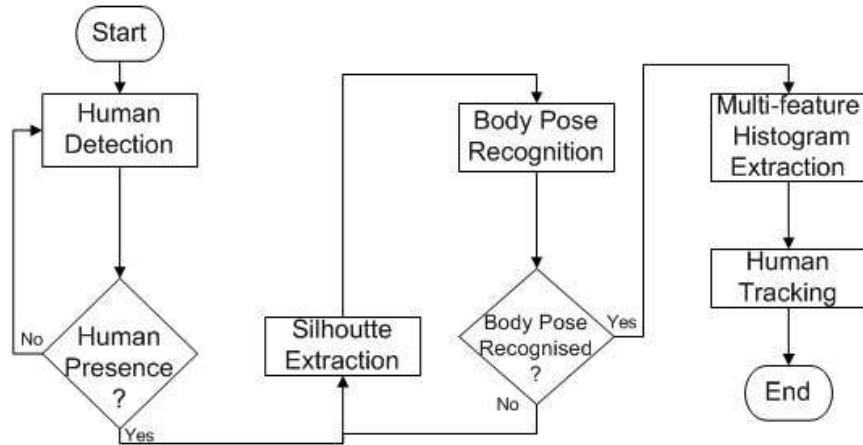


Figure 3: Flowchart of Human Tracking System

One of the possible approach to detect humans in cluttered street scenes is to use a statistical model. A set of training images containing a instance of the object can be used to develop a model to detect people. In statistical modeling multiple "positive" images and multiple "negative" (no instance of object) samples are used to train a classifier. Features are extracted from the training samples that can uniquely classify pedestrians from other moving objects in the scene. In this research, Haar-like features are extracted [51] and a large set of simple "weak" classifiers, that use a single feature to classify whether the image is human or non-human is used. The approach is similar to the technique developed by Viola and Jones [11] and later on extended by Lienhart et al. [51]. In the previous systems this approach was successfully used to detect faces. In this research we adapt that framework to detect pedestrians in thermal infrared spectrum.

Reviewing the literature we find there has been other statistical machine learning approaches for human detection. In the works like Papageorgiou et al. [52] have proposed a detector based on Haar wavelets and Support Vector

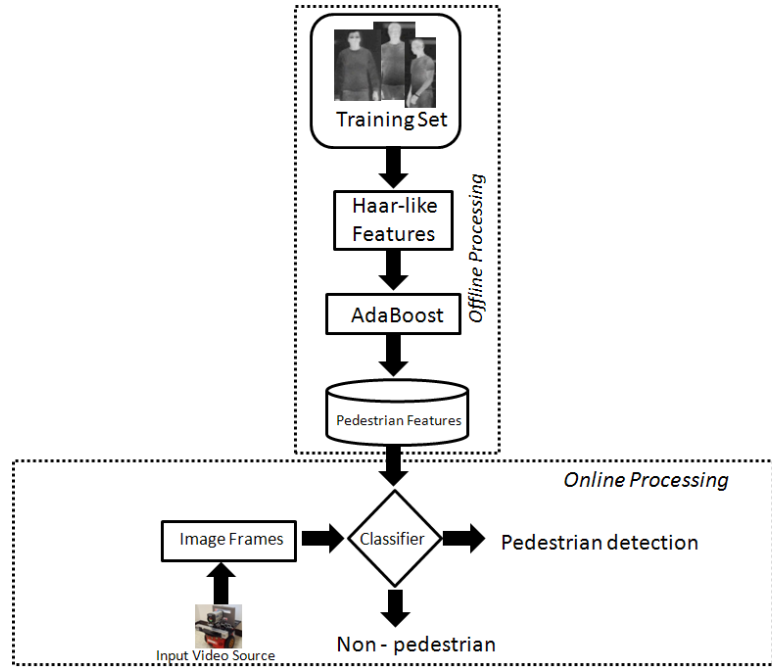


Figure 4: Flowchart of human detection module in thermal vision

Machines(SVM). Dalal and Triggs [53] proposed a method to combine Histogram of Oriented Gradient with support vector machine (SVM) for human detection. In this project a fast and reliable algorithm is required to detect humans in the thermal scene. Machine learning algorithm present good detection results at the cost of high computation. Inspired from the performance of face detection, a methodology to combine Haar-like features and AdaBoost algorithm is proposed to detect human objects in thermal vision.

3.1.1 Haar-like Features

The features extracted from each image in the training samples can be represented using a template (shape of the feature) and size of the feature (its scale). A subset of this features prototype is shown in figure 5. Each feature is composed of two or three "black" and "white" rectangles joined together. These

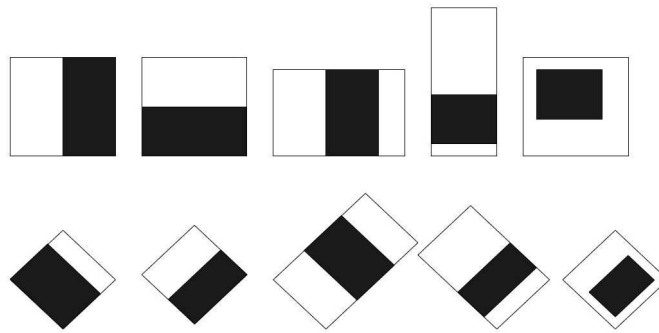
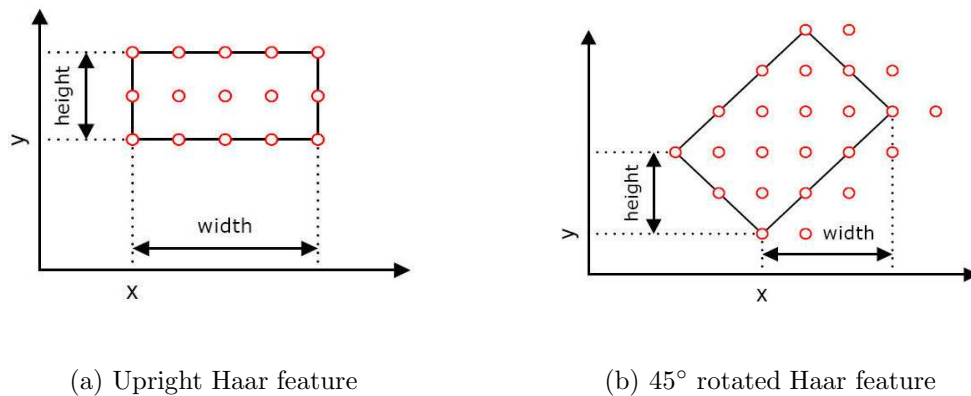


Figure 5: Subset of the Haar-like features in human detection



(a) Upright Haar feature

(b) 45° rotated Haar feature

Figure 6: Examples of a upright and rotated feature

rectangles can be up-right or rotated by an angle of 45 degrees. The value for the Haar-like features is calculated as a weighted sum of two components: the pixel gray level value summed over the whole feature area. The weights of these two components are of opposite signs and for normalization purpose, the absolute values are inversely proportional to the areas.

Computation of the pixel sums over multiple rectangles on the image directly would make detection task very slow and not suitable for real-time applications. Viola et al.[11] introduced an efficient method for computing the sums quickly. In that paper an integral image or Summed Area table (*SAT*) is computed over the

whole image I where SAT is defined as

$$SAT(x, y) = \sum_{i < x, j < y} I(i, j) \quad (19)$$

The pixel sum using rectangle corners coordinates is given by the following equation

$$S(r) = SAT(x_0+w, y_0+h) - SAT(x_0+w, y_0) - SAT(x_0, y_0+h) + SAT(x_0, y_0) \quad (20)$$

3.1.2 Classification Algorithm

Given a feature set and a training set of positive (people) and a negative (non-people) sample images, any number of machine learning approaches could be used to learn a classification function. In this research we use a AdaBoost algorithm for selecting a small set of features from the pool of features and train a weak classifier to give the best discrimination between object and non-object. AdaBoost learning algorithm is used to boost the classification performance of the weak classifiers. Weak classifiers are designed to select the single feature which best separates the positive and negative examples. For each feature, the weak learner determines the optimal threshold classification function, so that the minimum number of examples are misclassified. A cascade of such classifiers is constructed which achieves increased detection performance while radically reducing the computation time. The key objective is to construct smaller, and therefore more efficient, boosted classifiers, which reject many of the negative sub-windows, while detecting almost all positive instances. Simple classifiers are used to reject most sub-windows before more complex classifiers are called upon, in order to achieve low false positive rates. A cascade of classifiers is a degenerated decision tree where, at each stage, a classifier is trained to detect almost all objects of interest (pedestrians or other objects,) while rejecting a certain fraction of the non-object patterns.

Each stage was trained using the AdaBoost algorithm. AdaBoost is a powerful machine learning algorithm that can learn a strong classifier based on

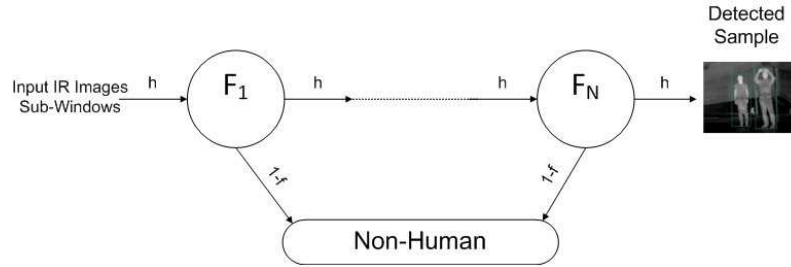


Figure 7: Cascade of boosted classifiers with N-stages

a (large) set of weak classifiers by re-weighting the training samples. The feature-based classifier that best classifies the weighted training samples is added at each round of boosting. As the stage number increases, the number of weak classifiers, needed to achieve the desired false alarm rate at the given hit rate, also increases. The cascade training process involves two types of trade-offs. In most cases, the classifiers with most features will achieve higher detection rates and lower false positive rates. At the same time, classifiers with more features require more time to compute. One could define an optimization framework in which: i) the number of classifier stages, ii) the number of features in each stage, and iii) the threshold of each stage, are traded off in order to minimize the expected number of evaluated features. Unfortunately, finding this optimum is a tremendously difficult task. In practice, a very simple framework is used to produce an effective classifier which is highly efficient. Each stage in the cascade reduces the false positive rate, as well as the detection rate. A target is selected for the minimum reduction in false positives and the maximum decrease in detection rate. Each stage is trained by adding features until the target detection and false positives rates are met (these rates are determined by testing the detector on a validation set). Stages are added until the overall target for false positive and detection rate is met. A total of 700 positive image samples were manually selected and 1500 background images

is labeled as negative images. These images were used to train the classifier. The detection of people in a thermal scene is done by sliding a search window through the frame image and checking whether an image region at a certain location is classified as pedestrian or non-pedestrian.



(a) True positive detection



(b) False negative detection



(c) False positive detection

Figure 8: Human detection results

3.2 Body Pose Recognition

In order to track moving human object and identify threat, a human pose recognition module is implemented to identify human with known body poses. The objective is to identify a human object with a known gesture in the observed scene among the detected humans from the previous stage of human detection. Once identification is successful the histogram of the human is extracted and the

target histogram is set which is used as target model for the tracking module to track that particular object in the future sequences. This functionality gives an added advantage of identifying and later on tracking humans with known threat postures. The functionality is designed to identify a certain body posture of the human object and detect the type of gesture. If the gesture is any known type of threat gesture the object is detected and subsequently the histogram of the object is extracted and stored for the tracking module to track using mean-shift algorithm.

The human silhouette which corresponds to the human body area in the thermal image is extracted by thresholding the thermal image and then the Hu-invariant moment is computed of the extracted contour. The Hu moments computed then is compared with the Hu-moment of a known silhouette which represents a particular gesture of the human body. Comparing moments of two contours is an very efficient way. Computing the moment of a contour is basically summing over all the pixels of the contour and can be defined as

$$m_{p,q} = \sum_{i=1}^n I(x, y)x^p y^q \quad (21)$$

where p is the x-order and q is the y-order.

The central moment can be computed the same way as the moments except the values of x and y used are displayed by the mean values:

$$\mu_{p,q} = \sum_{i=1}^n I(x, y)(x - x_{avg})^p (y - y_{avg})^q \quad (22)$$

The Hu-invariant moments are linear combinations of the normalized central moments given by the expression

$$\eta_{p,q} = \frac{\mu_{p,q}}{m_{00}^{\frac{p+q}{2}+1}} \quad (23)$$

By Combining the different normalized central moments it is possible to create an invariant function that is invariant to scale and rotation. A simple recognition



(a) “Y” Pose(P_1)



(b) “T” Pose(P_2)



(c) “Left Arm Raised” Pose(P_3)

Figure 9: Symmetrical body poses used for target identification

method for silhouette poses based on the moment descriptors is implemented. The seven higher-order moments provide excellent shape descriptors that are translation and scale invariant to the silhouettes with only minimal computation (requiring only a single pass of the image data). Since these moments are of different orders, a simple Euclidean metric cannot be used for matching. Accordingly, the Mahalanobis distance metric is used for matching based on a statistical measure of closeness to training examples. The distance measure is given by the following equation:

$$mahal(\mathbf{x}) = (\mathbf{x} - \mathbf{m})^T K^{-1} (\mathbf{x} - \mathbf{m}) \quad (24)$$

where \mathbf{x} is the moment feature vector, \mathbf{m} is the mean of the training moment vectors and K^{-1} is the inverse covariance matrix for the training vectors. This equation basically represents a hyper-ellipse with its center at \mathbf{m} and principal axes aligned with the eigenvectors from the covariance matrix K . This distance gives a variance measure of the input to the training class. With enough training vectors, a reliable statistical threshold can be set to announce whether or not the input vector is statistically close to a particular training class. However, to indicate the discriminatory power of these moment features for the silhouette poses, only a few examples of each pose are required (at least sufficient for matrix inversion) to relate the distances between the classes.

Table 2: Distance results for pose recognition

	P_1	P_2	P_3
T_1	14	204	2167
T_2	411	11	11085
T_3	2807	257	28

In this research, the training set consisted of 5 repetitions of 3 gestural poses (“Y”, “T”, and “Left Arm Raised”) shown in figure 9 done by each of five people. Table 2 shows the Mahalanobis distances for these new test poses T_i matched against the stored training models P_i . The table correctly shows that the true matches for the test poses (along the diagonal) have distances considerably smaller than to the other model poses in each row even though the first two poses, “Y” and “T”, are fairly close to one another. This is a typical result showing that thresholds tend to be easy to set with distances between even fairly similar gestures (“Y” and “T”) still about an order of magnitude apart.



Figure 10: Recognition of “T” pose in body pose recognition module

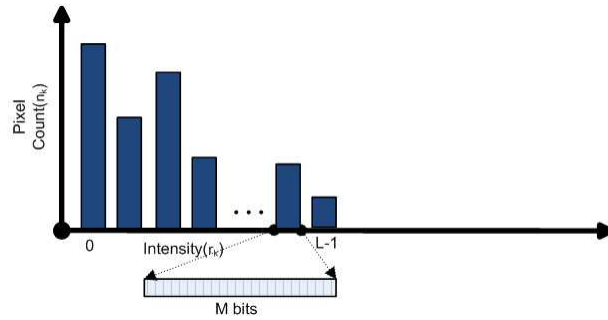
3.3 Multi-feature Histogram

The histogram of an image with intensity levels in the range of $[0, L - 1]$ is a discrete function of $h(r_k) = n_k$, where r_k is the k^{th} intensity value and n_k is the number of pixels in the image with intensity r_k . In digital computers each pixel of an image frame is represented using a certain number of bits depending upon the depth of the image. For example a M bit bitmap image means each pixel of that image is represented by M bits and therefore the total bin size of the image histogram is $2^M = L - 1$ bins. Each bin of the histogram represents a particular intensity value of the image and the image histogram shows the distribution map of the frequency of occurrence of the various pixels. Figure 11a shows the schematic diagram of a M bit intensity image histogram.

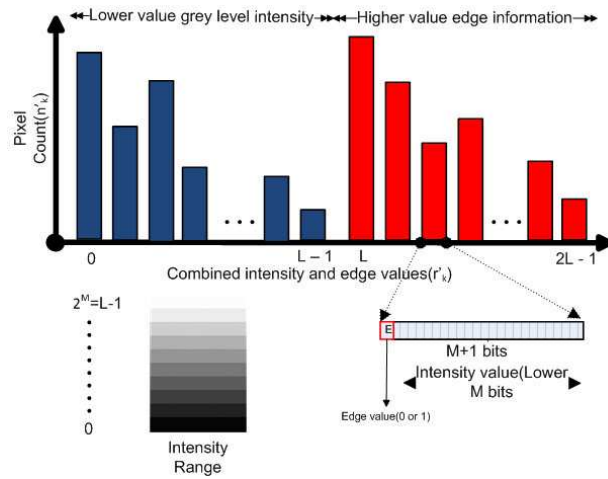
Although intensity pixel information used in image histogram is a very useful representation of an object in a scene it lacks shape information. In low contrast thermal IR imagery, objects in the foreground are represented using high brightness intensity and cold background is represented using low values of intensity making them dark. This representation is quite useful for locating

warm body objects in a scene, however has limitation in complex scenes where a objects with similar brightness intensity appears in the scene. Since the object representation completely lacks shape information with regard to the object of interest, it is easy for the tracker to drift away in the presence of other similar bright objects in the scene.

In order to increase the robustness of the mean shift tracking procedure in



(a) Schematic diagram of M bit intensity image histogram



(b) Schematic diagram of M+1 bit multi-feature image histogram

Figure 11: Schematic diagram Normal Vs Multi-feature histogram

thermal IR scenes the multi-feature histogram of the object is generated using two information. Along with the intensity information, the shape information in terms

of the object edge is also fused. The multi-feature histogram is defined as discrete function of $h(r'_k) = n'_k$, where r'_k is the k^{th} intensity value and n'_k is the number of features in the image with feature value r'_k which is represented using $M + 1$ bits. The total bin size of the multi-feature histogram is $2^{M+1} = 2L - 1$. The lower M bits represent the brightness value of the particular pixel and the last $M + 1^{th}$ bit in the MSB represents the edge information of that particular pixel. This $M + 1$ bit information regarding one pixel is the multi-feature information for a region of interest in the scene. The multi-feature histogram used to represent the object of interest contains both intensity as well as shape information. Figure 11b shows the schematic diagram of a multi-feature histogram. The multi-feature histogram has feature levels in the range of $[0, 2L - 1]$ of which $[0, L - 1]$ represents the brightness intensity information of the object without any edge information. The range $[L, 2L - 1]$ represents the area in the histogram where edge information is available of a particular pixel. Figure 12 compares the differences in normal and multi-feature histogram of two different human objects in a thermal scene.

Figure 12a shows a typical thermal scene inside the departmental lab corridor. Figure 12b and figure 12d are the two object of interests (human) with similar body temperatures but different edge distribution thus having similar gray level intensity distribution as seen in figure12f and figure 12g. Figure 12h and figure12i shows the multi-feature histogram of the two objects which is the combined distribution of the intensity as well as the edge information. The blue bars indicate the the gray level intensity distribution which is similar for both normal and multi-feature histograms. The red bars indicate the edge information of the objects and combined histograms show distinct distribution dissimilarities between the two objects.

The edge information defined here is based on the estimated edge direction.



(a) Thermal scene

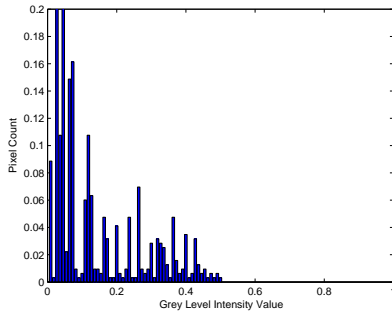


(b) Object 1

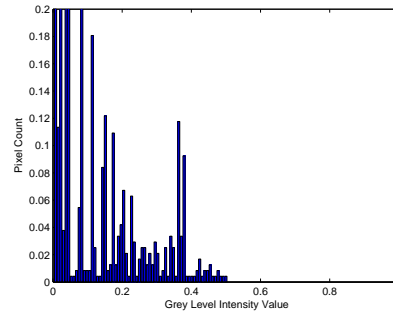
(c) Edge diagram 1

(d) Object 2

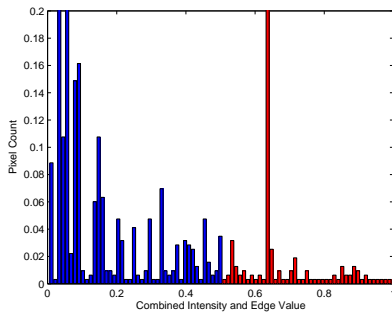
(e) Edge diagram 2



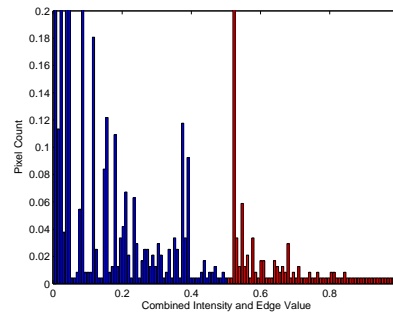
(f) Normal intensity histogram 1



(g) Normal intensity histogram 2



(h) Multi-feature histogram 1



(i) Multi-feature histogram 2

Figure 12: Normal vs Multi-feature histogram

Given an image region R corresponding to a target, the locations of the pixels in that region are $\{\mathbf{x}_i\}_{i=1,\dots,n}$ and the intensities of the pixels in that region are $I(R)$. The edge images are constructed by estimating the gradients $\frac{\partial I}{\partial x}$ and $\frac{\partial I}{\partial y}$ respectively, in the x and y directions by Sobel operators. The edge strength G and direction α at location \mathbf{x} are then approximated as

$$G(\mathbf{x}) = [(\frac{\partial I}{\partial x})^2 + (\frac{\partial I}{\partial y})^2]^{\frac{1}{2}} \quad (25)$$

$$\alpha(\mathbf{x}) = \tan^{-1}(\frac{\partial I}{\partial y} / \frac{\partial I}{\partial x}) \quad (26)$$

The edge direction is filtered to retain edges with magnitude above a predefined threshold (Th),

$$e(\mathbf{x}) = \begin{cases} \alpha(\mathbf{x}) & \text{if } G(\mathbf{x}) > \text{Th} \\ 0 & \text{otherwise} \end{cases} \quad (27)$$

The gray intensity and the edge information is combined to get a multi-feature histogram with a bin size of 2^{M+1} where each pixel of the feature component is represented using a $M + 1$ bit memory location. The first M bits from LSB indicate the gray intensity value and the MSB indicates the edge value. Tracking is performed by comparing the multi-feature histogram of the target model that was extracted from the body-pose recognition phase, with the target candidate and a similarity function is maximized between the the two probability distributions using Bhattacharyya coefficient [30].

Figure 13 shows a comparison study of the distribution of multi-feature histograms of three different human objects in a typical thermal scene. Figure 13a and figure 13b are two human objects having high and medium amount of edges as shown in figures 13d and 13e respectively. Figure ?? is another human object with less amount of edge information as shown in figure 13f. The percentage of edge feature's contribution to the computation of multi-feature histogram of the three objects are shown figures 13g,13h,13i respectively.



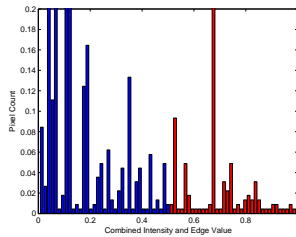
(a) Object 3

(b) Object 4

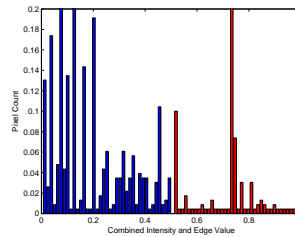
(c) Object 5



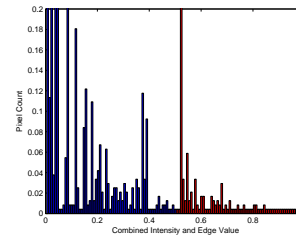
(d) Edge diagram 3 (e) Edge diagram 4 (f) Edge diagram 5



(g) 37.11%



(h) 33.37%



(i) 29.72%

Figure 13: Different edge contributions to multi-feature histograms of objects

3.4 Human tracking via Multi-feature Histogram

In case of infrared target tracking problems, the gray cue has gained the most attention, as it is well distinguished from the infrared target. However, it is vulnerable to the disturbance of similar gray background regions. The proposed multi-feature histogram enhances the tracking robustness in the mean-shift tracking framework. It takes advantages of the edge information to assist the gray level intensity information to model the structure of the human body.

Gray is a salience target feature for infrared target tracking. In the infrared

imaging system, the radiation heat from a human body is generally higher than the static background, such as a house or a road. Thus the human body often possesses higher gray values in the infrared image than the background, which can be used for distinguishing the target from background. To use the mean shift algorithm, a gray intensity probability density function indicating the target's gray region (extracted from the pose recognition module) is first computed by the histogram back-projection: the gray level intensity histogram of the target is calculated and stored in a lookup table. When a new frame comes in, the table is looked up for each pixel's gray, and a gray probability value is assigned to each pixel. Hence a probabilistic distribution map is obtained. The mean shift procedure can be employed (as discussed in chapter 2) to find the nearby dominant distribution peak.

Although the gray cue is useful for infrared target tracking, its main drawback lies in the disturbance of other similar gray regions, which may result in errors in the tracking process. To enhance the tracking robustness, the use of multi-feature histogram is proposed in this thesis which takes advantage of the edge information. The combined gray level intensity and edge information of the target model are described by a multi-feature probability density function that is calculated using the multi-feature histogram.

The task of finding the target location in the current frame using the Bhattacharyya coefficient can be formulated in the following way. If we assume that \mathbf{z} represents the multi-feature of the target model which is the $M + 1$ bit pixel information containing both gray level intensity as well as edge information, then feature \mathbf{z} is assumed to have a density function $q_{\mathbf{z}}$, while the candidate centered at a location \mathbf{y} has the feature distributed according to $p_{\mathbf{z}}(\mathbf{y})$. The problem of target tracking is then reduced to find the discrete location \mathbf{y} whose associated

density function $p_{\mathbf{z}}(\mathbf{y})$ is most similar to the target density $q_{\mathbf{z}}$.

A Bhattacharyya coefficient can be defined by

$$\rho[p(\mathbf{y}), q] = \int \sqrt{p_{\mathbf{z}}(\mathbf{y})q_{\mathbf{z}}} d\mathbf{z} \quad (28)$$

The derivation of the Bhattacharyya coefficient from the sample data involves the estimation of the density function p and q , for which a histogram formulation is carried out. The discrete density $\hat{\mathbf{q}} = \{\hat{q}_u\}_{u=1\dots M+1}$ is estimated where $\sum_{u=1}^{M+1} \hat{q}_u = 1$ from a $M + 1$ bin multi-feature histogram of the target model, while $\hat{\mathbf{p}}(\mathbf{y}) = \{\hat{p}_u(\mathbf{y})\}_{u=1\dots M+1}$ where $\sum_{u=1}^{M+1} \hat{p}_u = 1$ is estimated at a given location \mathbf{y} from the $M + 1$ bin multi-feature histogram of the target candidate. Therefore, the sample estimate of the Bhattacharyya coefficient is given by

$$\rho[\hat{\mathbf{p}}(\mathbf{y}), \hat{\mathbf{q}}] = \sum_{u=1}^{M+1} \sqrt{\hat{p}_u(\mathbf{y})\hat{q}_u} \quad (29)$$

The geometric interpretation of equation 29 is the cosine of the angle between the $M + 1$ dimensional unit vectors $(\sqrt{\hat{p}_1}, \dots, \sqrt{\hat{p}_{M+1}})^T$ and $(\sqrt{\hat{q}_1}, \dots, \sqrt{\hat{q}_{M+1}})^T$. Using 29 the distance between the two distributions is defined as

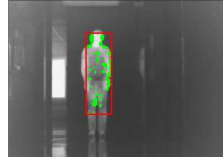
$$d(\mathbf{y}) = \sqrt{1 - \rho[\hat{\mathbf{p}}(\mathbf{y}), \hat{\mathbf{q}}]} \quad (30)$$

The tracking procedure is thus the minimization of 30 or maximization of the Bhattacharyya coefficient which is achieved using mean shift iterations as described previously in chapter 2.

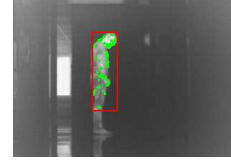
In figure 14 the multi-feature histograms of a few frames in a tracking sequence is shown. As explained the tracking procedure is basically the minimization of equation 30 or maximization of Bhattacharyya similarity function (equation 29) between the target model and the target candidate between every two frames of a thermal video sequences. The minimum distance between the target model (histogram of previous frame) and the target candidate (histogram of current



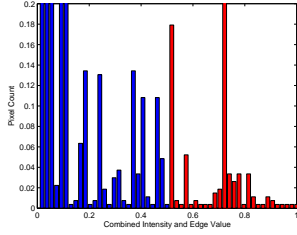
(a) Frame No. 100



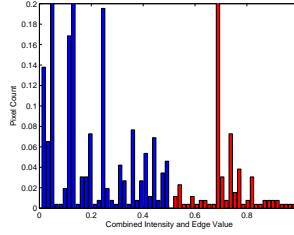
(b) Frame No. 280



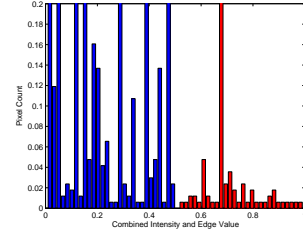
(c) Frame No. 460



(d) $d(\mathbf{y}) = 0.334$



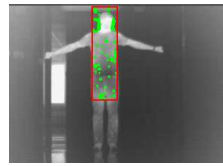
(e) $d(\mathbf{y}) = 0.372$



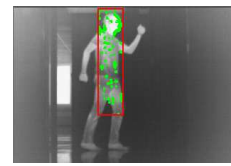
(f) $d(\mathbf{y}) = 0.351$



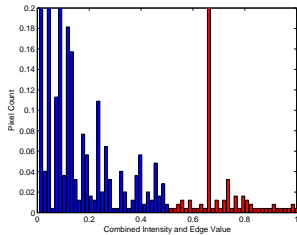
(g) Frame No. 640



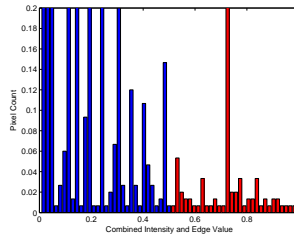
(h) Frame No. 820



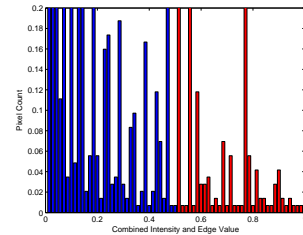
(i) Frame No. 970



(j) $d(\mathbf{y}) = 0.368$



(k) $d(\mathbf{y}) = 0.387$

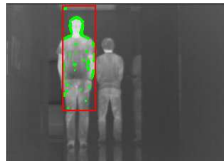


(l) $d(\mathbf{y}) = 0.424$

Figure 14: Multi-feature histograms in a tracking sequence

frame) is also shown in the figures. In figure 15 the multi-feature histogram of the object of interest (human object) is shown in a more complex tracking scenario. In this figure, tracking is carried out in the presence of another similar moving object (another human being)

The detailed procedure of the whole human tracking system has been



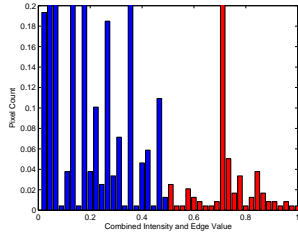
(a) Frame No. 55



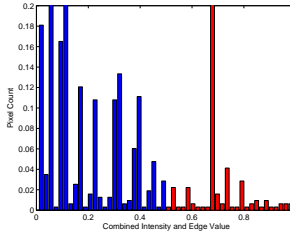
(b) Frame No. 150



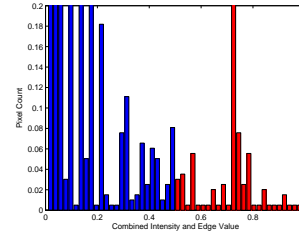
(c) Frame No. 300



(d) $d(\mathbf{y}) = 0.224$



(e) $d(\mathbf{y}) = 0.315$



(f) $d(\mathbf{y}) = 0.499$

Figure 15: Multi-feature histograms in thermal tracking sequence

discussed in this chapter. The flowchart of the whole system presented shows how the control of the program is passed from one module to the next one. The multi-feature histogram is introduced and various instances of the multi-feature histogram is shown in different tracking scenarios. In the next chapter the results form each module is discussed along with experimental results and performance comparisons with the current state-of-the-art in human tracking system.

CHAPTER 4

PERFORMANCE EVALUATION

This chapter presents the set-up used to conduct the experimental part of the thesis. The mobile robot platform - a Pioneer P3-DX mobile robot is first introduced. The human tracking system described in this thesis was entirely implemented on this robotic platform. The characteristics of the visual sensor used in this research is also discussed. The process of collecting the ground truth data is presented. The metrics used for evaluation of the tracking system is explained. This chapter also introduces the benchmark dataset used for performance evaluation. The results are discussed and tracking performance is compared to standard mean shift algorithm.

4.1 Experimental Setup

The experimental setup used in this thesis is a Pioneer P3-DX mobile robot, which is a research mobile robot platform especially designed for all terrain (outdoor and indoor) navigation. The base of the Pioneer P3 DX platform is fully assembled with motors with 500-tick encoders, 19cm wheels, tough aluminum body and 8 forward-facing ultrasonic (sonar) sensors. The base Pioneer P3 DX platform can reach speeds of 1.6 meters per second and carry a payload of up to 23 kg. A thermal Forward Looking Infrared Camera (FLIR Thermovision A40) is mounted on the mobile robot platform. Information captured through the thermal infrared camera is used for people detection and tracking. The system is also equipped with a Intel Core 2 Duo (2 GHz) processor laptop computer with both Windows and Linux operating system. The code for the tracker software is implemented in C++ using Microsoft Visual Studio IDE. Figure 16 shows the experimental mobile

robot setup used for testing the proposed human tracking procedure.

The camera is connected to the computer using a frame grabber which allows capturing of data at a rate of 30 frames per second. The thermal camera converts infrared radiation into an image where each pixel corresponds to a temperature value. In this set-up the visible range in the grey-scale image is equivalent to the temperature range from 75 to 98.6 degree Fahrenheit.

The human tracking system was tested by deploying the robot during several

Thermal Sensor : FLIR Thermovision A40



Mobile Robot platform : Pioneer P3-DX

Figure 16: Mobile robot setup

runs. The robot was operated in both indoor (a corridor and lab room at our institute) and outdoor (courtyard in front of the College of Engineering building) environments. People taking part in the experiments were asked to walk in front of the robot while it performed different autonomous scanning behaviors while the robot was stationary. At the same time, image data were collected with a frequency of 30 Hz. The resolution of the thermal images was 320×240 pixels. The robot starts first in a search mode, scanning continuously at the same time trying to detect a person in the thermal image. After a person is detected the robot tries to get closer and maintain a constant distance to the person. This is realized within an image-based control loop: the direction of the robot is adjusted

so that the position of the person provided by the tracker remains in the middle of the thermal image. The velocity of the robot is determined by the height of the person which corresponds to the apparent distance between the robot and person. If the height is bigger than a specified threshold, or in other words the robot gets close enough to a person, the robot stops. Any change in the position of a person is immediately compensated by the control loop resulting in an appropriate action of the robot. The high frame rate of the system allowed for smooth operation in both indoor and outdoor environments.

4.2 Evaluation of Tracking System

To evaluate the effectiveness of the proposed multi-feature histogram based mean-shift tracking algorithm, the proposed method was tested and compared with the traditional mean shift tracking algorithm against several test sequences. The test video sequences were obtained from the Object Tracking and Classification in and beyond the Visible Spectrum (OTCBVS) benchmark dataset. These video datasets were captured under the condition of fixed camera placed on top of a three stored building in the campus of Ohio State University (OSU). The videos show a busy pathway on the campus of OSU. The image frames size was 320×240 pixel. The proposed tracking method was tested on the benchmark dataset and the quantitative analysis of three different conditions are shown. Along with the benchmark dataset the proposed algorithm was also tested on video sequences obtained through the robotic setup explained in the previous section.

Obtaining the ground truth in the case of video data is often a difficult, monotonous and labor demanding process. Optimally the ground truth data should consist of true values for each component of the state vector. Then the errors for each state variable could be specified to obtain an indicator of the performance of the tracker. In this thesis the centroid of the bounding box area around a person

was considered in order to simplify the ground truth labelling process. The top and bottom edges of a bounding box were determined from the contours of the head and feet while the sides were specified by the maximum width of the torso (without arms). The centre of the bounding box is determined and the position coordinate is taken as the ground truth value for the position of the object. The cases when persons appeared too close ($< 3\text{m}$) or too far ($> 10\text{m}$) to the robot were not taken into account. Another measurement taken into consideration was the height and the width of the bounding box around the object of interest. Therefore, the ground truth state vector for each frame can be denoted by $T = (x, y, h, w)^T$ where (x, y) is the centre of the bounding box and h and w are the true value of the height and width of the bounding box surrounding the object. This type of ground truth information is just an approximation, and the quality of this process is affected by factors such as the naturally blurred appearance of a person in the image, noise caused by the movement of the robot and also the skill of the person labeling the data.

The output from the tracking system is a set of tracks corresponding to the person being tracked. A single track is a collection of estimated values of the state vector for the corresponding person over some period of time. To compare the output from the tracker with the ground truth data, we first transform the information provided by the mean shift to match the assumed ground truth data described in the previous paragraph. Bounding boxes from the ground truth data are referred to as targets and those from the tracker as candidates. The output of the tracker is a state vector $E = (x', y', h', w')$ where (x', y') is the estimated centre of the tracker and h' and w' is the estimated height and width of the tracker bounding box respectively. The output from the tracker should produce results as close to ground truth as possible. However, due to tracking errors, tracks deviate

from their true paths, get scattered, missing or swapped. These phenomena occur especially in the case of tracking a single person in a crowded scene.

In order to calculate the tracking error, the Euclidean distance between the state vectors T and E is calculated using the following equation

$$error = \|T - E\| = \sqrt{(x - x')^2 + (y - y')^2 + (h - h')^2 + (w - w')^2} \quad (31)$$

4.3 Quantitative Results

The quantitative analysis of the proposed tracking system is presented in this section. The tracker was evaluated on the OTCBVS benchmark dataset and the results was compared with that of the traditional mean shift tracking algorithm. The three analytic results shown represents the tracker performance on benchmark video dataset having three different characteristics in terms of external conditions. Table 3 show the different characteristics of the video sequences.

Table 3: Characteristics of video sequences used for evaluation

Sequence Type	Number of sequences	Sequence Characteristics	Target size
S1	8	Background disturbances	18×10
S2	8	Illumination changes	13×7
S3	5	Similar Objects	18×7

Figure 17 show snapshots of the three types of thermal data sequences on which the proposed multi-feature histogram was implemented for performance evaluation. Figure 17a is an example where an object with similar brightness intensity as that of the object being tracked is present in the scene. Figure 17b is a thermal data sequence with illumination variation and figure 17c is tracking sequence where two similar looking objects cross each other in the thermal scene from opposite directions. The performance evaluation of the three mentioned tracking



(a) Complex background (S1)



(b) Illumination variations (S2)

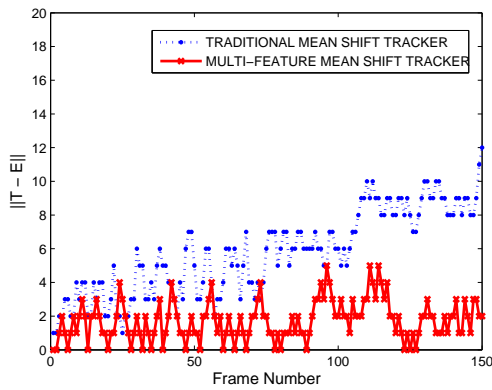


(c) Similar Objects (S3)

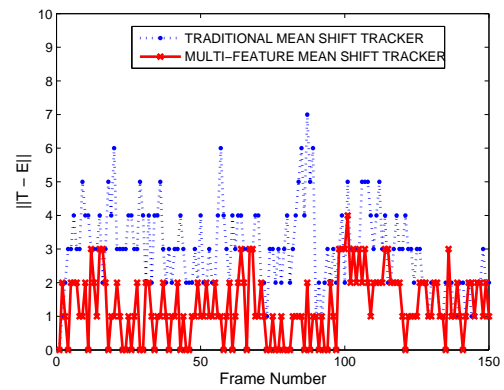
Figure 17: Examples from OTCBVS dataset

sequences are shown in figure 18. Figure 18a shows the plot of tracking error when the multi-feature histogram mean shift tracker was implemented on S1 dataset. The traditional mean-shift algorithm was also implemented on this dataset and the tracking error was plotted for comparison. The results show the failure of traditional mean-shift algorithm due to the presence of disturbance in the form of a vehicle in the background. The error plot of the multi-feature meanshift tracker show robustness in the presence of disturbance. Figure 18b shows the tracking error plots on sequence S2 where the illumination of the scene changes over time. Figure 18c is the tracking error plots when two similar looking human objects are present in the scene. An instance of partial occlusion also takes place in this data sequence. The traditional mean-shift tracker loses track of the

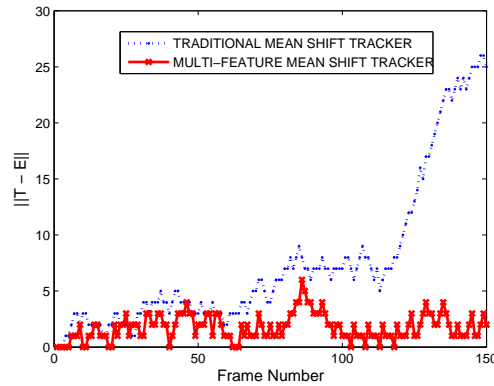
correct object after two moving objects cross each other whereas the multi-feature histogram does not lose the tracks of the object of interest. Table 4 shows the mean absolute error computed on the OTCBVS benchmark dataset. The overall error indicates improvement in the performance of the multi-feature histogram tracker when compared to traditional mean shift tracking algorithm.



(a) Tracking error on sequence S1



(b) Tracking error on sequence S2



(c) Tracking error on sequence S3

Figure 18: Tracking error plots on OTCBVS datasets

Table 4: Mean Absolute error(MAE) on OTCBVS benchmark dataset

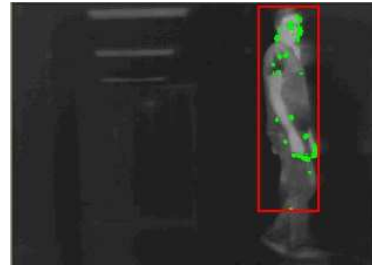
Sequence Type	Traditional Mean Shift tracker	Multi-feature mean shift tracker
S1	3.6664	1.9745
S2	3.4524	1.8606
S3	5.0062	2.0161
Overall	3.9039	1.9305

4.4 Experimental Results

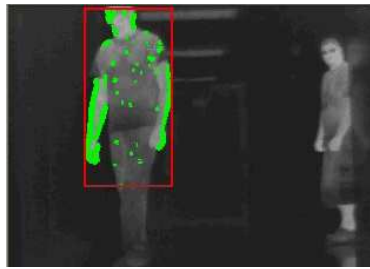
In this section the qualitative results for the proposed multi-feature based mean shift tracking algorithm implemented using a Pioneer P3-DX mobile robot platform is discussed. The system was tested by performing several test runs on actual human objects in various tracking scenarios. Figure 19 is a set of frames taken from a tracking run performed inside the corridors of the department of electrical and computer engineering at Temple University. The example shows tracking of a single human object in presence of similar moving human objects in the same thermal scene. Figure 20 is another tracking example showing tracking of a single person in the presence of another similar looking human object in the thermal scene. These above results shows that tracking is robust in the presence of similar looking moving objects and the problem of shifting of tracker in the thermal scene has been avoided using the multi-feature histogram in the mean shift tracking framework. Figure 21 is a a output of a tracking system in a outdoor environment. In this image sequences the tracker shows robust tracking capability in uncontrolled outdoor and tracking capability is presented in presence of limited duration partial occlusion. Figure 22 shows tracking is robust under different poses of human body orientation once the object of interest has been locked for tracking. In figure 23,24 the actual tracking of an human object in an outdoor condition is shown. The images sequences shows the experimental set-up discussed in the beginning of the chapter, tracks an object of interest in a tracking scenario. Tracking is robust



(a) Frame No. 110



(b) Frame No. 410



(c) Frame No. 710



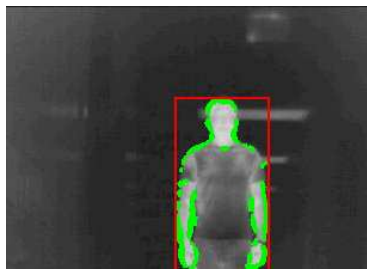
(d) Frame No. 1010



(e) Frame No. 1310

Figure 19: Tracking results in indoor environment(camera view)

under variable lighting conditions, cluttered background and in presence of partial occlusion which are some natural events possible in a real tracking scenario.

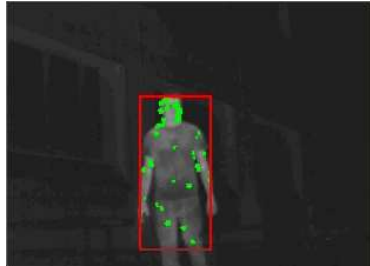


(a) Frame No. 410

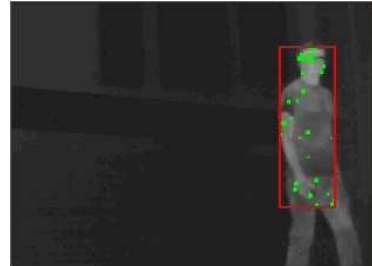


(b) Frame No. 710

Figure 20: Tracking of a single person



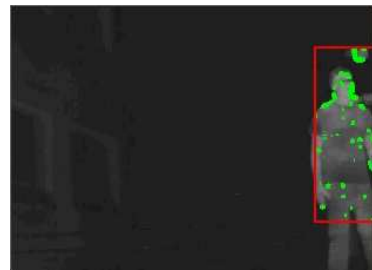
(a) Frame No. 110



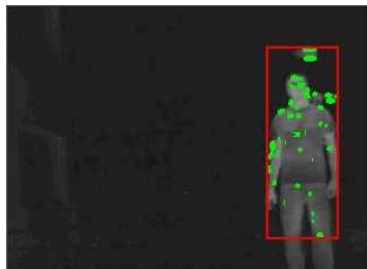
(b) Frame No. 410



(c) Frame No. 710



(d) Frame No. 1010



(e) Frame No. 1310

Figure 21: Tracking results in outdoor environment(camera view)



(a) Frame No.840



(b) Frame no. 1240

Figure 22: Robust tracking under different pose orientations



(a)



(b)

Figure 23: Tracking in outdoor conditions



(a)



(b)



(c)



(d)



(e)

Figure 24: Tracking in the presence of multiple moving objects

CHAPTER 5

CONCLUSION

In this thesis a human tracking system based on multi-feature histogram is proposed that can autonomously detect, identify and track a human object in a complex thermal scene. The scope of this research is in the development of the detection, identification and tracking algorithms to track moving human objects using a mobile robot platform. The mobile robot platform is an open source research platform used for various computer vision and robotics research.

A multi-feature histogram has been proposed to model the object region for successful tracking of a moving human object in the mean-shift tracking framework. The use of multi-feature histogram improves the robustness of the tracker and quantitative evaluation shows improvement in tracking performance in thermal scenes compared to traditional mean shift algorithm which uses a single feature histogram for object representation. The experimental results show robust tracking capability under varying condition of outdoor and indoor environments. The tracking is also near real time however due to limited viewing angle of the thermal IR camera the robot speed is drastically reduced in order to synchronize the robot movement with the tracking window. The whole setup consists of a ThermoVision A40 thermal IR camera and a ordinary laptop computer mounted on a Pioneer P3-DX Mobile Robot Platform.

The human tracking system has been divided into three major modules - human detection, body pose recognition and multi-feature histogram based tracking. The first two modules is essential for the successful detection of the object of interest to be tracked via the tracking module. In the human detection module the existing Adaboost classification algorithm has been adapted to detect

human objects in complex thermal scenes. After successful human object detection in the thermal scene the system control is transferred to a body pose recognition module where a simple shape matching criterion is checked using hu-invariant moments to identify the object of interest.

However scope of improvement is there to increase the precision of detection and tracking. In the human detection phase the accuracy rate of detection of people is directly proportional to the number of training images. In the human body posture recognition phase we have used a simple method to do shape analysis and matching to extract histogram of object of interest. However the real objective in this phase is to identify and recognize threat gestures in order to identify criminal activity. In the tracking phase of the system a modified mean shift tracking is used to track the object of interest. Mean Shift Tracking is an efficient technique for tracking 2D blobs through an image. It has gained much popularity for its simplicity and robustness. In this research a edge information is used as an extra cue to track a moving human in thermal vision which gives a even more robust mean shift tracker. The method is implemented and tested with real object tracking scenarios in indoor and outdoor environments. This research emphasizes on the detection and tracking of moving human objects keeping in mind the possibility of providing added support to law enforcement organizations in the field of surveillance.

In this research the whole method of tracking is an online process and only the training phase for people detection is the off-line computation in this system. The limitation of the system is the object should not be running or be out of sight of the camera for too long. Although the object tracking is capable of partial occlusion handling, the object tracker will tend to fail under total occlusion for a longer period of time. The speed of the robot has to be reduced sufficiently in

order to synchronize with the tracking window as the viewing angle of the thermal camera is quite small. With a larger viewing angle lens the robot speed can be increased for it to track quicker moving objects. A number of tracking examples has been shown in this thesis.

5.1 Future Work

The proposed human tracking system in this thesis is part of a much larger research project where the main objective is thermal analysis for threat detection in civilian surveillance scenarios. This system can be extended towards developing a mobile surveillance system that can autonomously scan a scene for threat detection. The implementation of this system on a mobile robot platform, shows the capability of tracking a human object in both indoor and outdoor environments. However, in order to realize a complete mobile surveillance solution few improvements are necessary.

In the body pose recognition module, a simple shape matching criteria using Hu-invariant moment, is used to match the shape of the object for symmetrical pose recognition. However, in a more realistic scenario, the objective of this module is to identify threatening gestures or estimate threat from the given set of detected human objects in the scene. More research in human motion analysis is required to classify properly human motions into threat and non-threat gestures.

Another possible area of improvement is to use both thermal and visible range camera together and use the fused visible and thermal images for human tracking. It should be interesting to study whether the fusion of the two types of input gives more precision in tracking scenarios. Combination of thermal and visible images has been successful in face recognition systems since both visible and thermal images provide complementary information which has improved face recognitions results. Since thermal images have very low resolution, it lacks a lot

of information required for successful human body pose recognition. On the other hand, visible images contain lot of information so can be used for human motion analysis. However, illumination variations is a major constraint in visible spectrum which can be dealt with thermal vision. Therefore a perspective extension of this research is the fusion of thermal and visible camera for better object recognition.

LIST OF REFERENCES

- [1] A. Yilmaz, O. Javed, and M. Shah, “Object tracking: A survey,” *ACM Comput. Surv.*, vol. 38, no. 4, p. 13, 2006.
- [2] Z. Chen and H.-J. Lee, “Knowledge-guided visual perception of 3-d human gait from a single image sequence,” *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 22, no. 2, pp. 336 –342, mar/apr 1992.
- [3] K. Rohr, “Towards model-based recognition of human movements in image sequences,” *CVGIP: Image Understanding*, vol. 59, no. 1, pp. 94 – 115, 1994. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1049966084710060>
- [4] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, “Pfinder: real-time tracking of the human body,” in *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, oct 1996, pp. 51 –56.
- [5] I. Haritaoglu, D. Harwood, and L. Davis, “W4: Who? when? where? what? a real time system for detecting and tracking people,” in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, apr 1998, pp. 222 –227.
- [6] K. Rohr, “Towards model-based recognition of human movements in image sequences,” *CVGIP: Image Underst.*, vol. 59, pp. 94–115, January 1994. [Online]. Available: <http://portal.acm.org/citation.cfm?id=183817.183826>
- [7] S. Niyogi and E. Adelson, “Analyzing gait with spatiotemporal surfaces,” in *Motion of Non-Rigid and Articulated Objects, 1994., Proceedings of the 1994 IEEE Workshop on*, nov 1994, pp. 64 –69.
- [8] A. Lipton, H. Fujiyoshi, and R. Patil, “Moving target classification and tracking from real-time video,” in *Applications of Computer Vision, 1998. WACV '98. Proceedings., Fourth IEEE Workshop on*, oct 1998, pp. 8 –14.
- [9] L. Brethes, P. Menezes, F. Lerasle, and J. Hayet, “Face tracking and hand gesture recognition for human-robot interaction,” in *Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on*, vol. 2, 26-may 1, 2004, pp. 1901 – 1906 Vol.2.
- [10] M. Turk and A. Pentland, “Face recognition using eigenfaces,” in *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR '91., IEEE Computer Society Conference on*, jun 1991, pp. 586 –591.

- [11] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, 2001, pp. I-511 – I-518 vol.1.
- [12] L. Zhao and C. Thorpe, “Stereo- and neural network-based pedestrian detection,” in *Intelligent Transportation Systems, 1999. Proceedings. 1999 IEEE/IEEEJ/JSAI International Conference on*, 1999, pp. 298 –303.
- [13] H. Nanda and L. Davis, “Probabilistic template based pedestrian detection in infrared videos,” in *Intelligent Vehicle Symposium, 2002. IEEE*, vol. 1, june 2002, pp. 15 – 20 vol.1.
- [14] E. Huber and D. Kortenkamp, “Using stereo vision to pursue moving agents with a mobile robot,” in *Robotics and Automation, 1995. Proceedings., 1995 IEEE International Conference on*, vol. 3, may 1995, pp. 2340 –2346 vol.3.
- [15] R. Kahn, M. Swain, P. Prokopowicz, and R. Firby, “Gesture recognition using the perseus architecture,” in *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR '96, 1996 IEEE Computer Society Conference on*, jun 1996, pp. 734 –741.
- [16] M. Bertozzi, A. Broggi, P. Grisleri, T. Graf, and M. Meinecke, “Pedestrian detection in infrared images,” in *Intelligent Vehicles Symposium, 2003. Proceedings. IEEE*, june 2003, pp. 662 – 667.
- [17] U. Meis, W. Ritter, and H. Neumann, “Detection and classification of obstacles in night vision traffic scenes based on infrared imagery,” in *Intelligent Transportation Systems, 2003. Proceedings. 2003 IEEE*, vol. 2, oct. 2003, pp. 1140 – 1144 vol.2.
- [18] F. Xu, X. Liu, and K. Fujimura, “Pedestrian detection and tracking with night vision,” *Intelligent Transportation Systems, IEEE Transactions on*, vol. 6, no. 1, pp. 63 – 71, march 2005.
- [19] D. Fox, W. Burgard, and S. Thrun, “Markov localization for mobile robots in dynamic environments,” *Journal of Artificial Intelligence Research*, vol. 11, 1999.
- [20] D. Schulz, W. Burgard, D. Fox, and A. Cremers, “Tracking multiple moving objects with a mobile robot,” in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, 2001, pp. I-371 – I-377 vol.1.
- [21] B. Kluge, C. Kohler, and E. Prassler, “Fast and robust tracking of multiple moving objects with a laser range finder,” in *Robotics and Automation, 2001. Proceedings 2001 ICRA. IEEE International Conference on*, vol. 2, 2001, pp. 1683 – 1688 vol.2.

- [22] S. Feyrer and A. Zell, *Robust real-time pursuit of persons with a mobile robot using multisensor fusion*. I O S PRESS, 2000, pp. 710–715.
- [23] S. Sun, D. Haynor, and Y. Kim, “Semiautomatic video object segmentation using vsnakes,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 1, pp. 75 – 82, Jan. 2003.
- [24] N. Paragios and R. Deriche, “Geodesic active regions for motion estimation and tracking,” in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 1, 1999, pp. 688 –694 vol.1.
- [25] A. Yilmaz, X. Li, and M. Shah, “Contour-based object tracking with occlusion handling in video acquired using mobile cameras,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 11, pp. 1531 –1536, nov 2004.
- [26] M. Isard and A. Blake, “Condensation – conditional density propagation for visual tracking,” 1998.
- [27] K. Fukunaga and L. Hostetler, “The estimation of the gradient of a density function, with applications in pattern recognition,” *Information Theory, IEEE Transactions on*, vol. 21, no. 1, pp. 32 – 40, Jan. 1975.
- [28] Y. Cheng, “Mean shift, mode seeking, and clustering,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 17, no. 8, pp. 790 –799, Aug. 1995.
- [29] D. Comaniciu and P. Meer, “Mean shift: A robust approach toward feature space analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, pp. 603–619, May 2002. [Online]. Available: <http://portal.acm.org/citation.cfm?id=513073.513076>
- [30] D. Comaniciu, V. Ramesh, and P. Meer, “Real-time tracking of non-rigid objects using mean shift,” in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, vol. 2, 2000, pp. 142 –149 vol.2.
- [31] E. Parzen, “On estimation of a probability density function and mode. the annals of mathematical statistics,” pp. 1065–1076, 1962.
- [32] D. Comaniciu, V. Ramesh, and P. Meer, “Kernel-based object tracking,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, pp. 564–575, May 2003.
- [33] R. Collins, “Mean-shift blob tracking through scale space,” in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 2, june 2003, pp. II – 234–40 vol.2.
- [34] L. L. H. Zhang and Q. Yu., “A scale rotation adaptive new mean shift tracking method,” p. 68330S, 2008.

- [35] J. Drarni and S. Roy, “A simple oriented mean-shift algorithm for tracking,” in *Image Analysis and Recognition*, ser. Lecture Notes in Computer Science, M. Kamel and A. Campilho, Eds. Springer Berlin / Heidelberg, 2007, vol. 4633, pp. 558–568.
- [36] A. P. Leung and S. Gong, “Mean shift tracking with random sampling,” in *Proc. BMVC 2005*, 2006, pp. 729–738.
- [37] G. R. Bradski, “Computer vision face tracking for use in a perceptual user interface,” 1998.
- [38] R. Duraiswami and L. Davis, “Efficient Mean-Shift Tracking via a New Similarity Measure,” *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, pp. 176–183, 2005.
- [39] D. Comaniciu and V. Ramesh, “Mean shift and optimal prediction for efficient object tracking,” in *Image Processing, 2000. Proceedings. 2000 International Conference on*, vol. 3, 2000, pp. 70–73 vol.3.
- [40] N. S. Peng, J. Yang, and Z. Liu, “Mean shift blob tracking with kernel histogram filtering and hypothesis testing,” *Pattern Recogn. Lett.*, vol. 26, pp. 605–614, April 2005. [Online]. Available: <http://dx.doi.org/10.1016/j.patrec.2004.08.023>
- [41] T. Lindeberg, “Feature detection with automatic scale selection,” *Int. J. Comput. Vision*, vol. 30, pp. 79–116, November 1998. [Online]. Available: <http://portal.acm.org/citation.cfm?id=305297.305298>
- [42] K. Smith, D. Gatica-Perez, and J.-M. Odobez, “Using particles to track varying numbers of interacting people,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, june 2005, pp. 962 – 969 vol. 1.
- [43] D. M. Gavrilu and L. S. Davis, “Towards 3-d model-based tracking and recognition of human movement: a multi-view approach,” *Int Workshop on Face and Gesture Recognition*, pp. 3–8, 1995. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.56.5329&rep=rep1&type=pdf>
- [44] H. Sidenbladh, “Probabilistic tracking and reconstruction of 3D human motion in monocular video sequences,” Numerical Analysis and Computer Science, Doctoral Dissertation, ISRN KTH/NA/P-01/14-SE, Oct. 2001.
- [45] I. Horswill, “Polly: A vision-based artificial agent.” in *AAAI’93*, 1993, pp. 824–829.
- [46] M. R. Blackburn and H. G. Nguyen, “Autonomous visual control of a mobile robot,” in *Proceedings of the 1994 Image Understanding Workshop*, 1994, pp. 1143–1150.

- [47] D. Kortenkamp, E. Huber, R. P. Bonasso, and M. Inc, “Recognizing and interpreting gestures on a mobile robot,” in *In Proceedings of AAAI-96*. AAAI Press/The MIT Press, 1996, pp. 915–921.
- [48] C. Schlegel, J. Illmann, H. Jaberg, M. Schuster, and R. Worz, “Vision based person tracking with a mobile robot,” in *In Proc. British Machine Vision Conf*, 1998, pp. 418–427.
- [49] S. Waldherr, S. Thrun, R. Romero, and D. Margaritis, “Template-based recognition of pose and motion gestures on a mobile robot,” in *In Proceedings of the AAAI Fifteenth National Conference on Artificial Intelligence*. MIT Press, 1998, pp. 977–982.
- [50] M. Kleinhagenbrock, S. Lang, J. Fritsch, F. Lomker, G. Fink, and G. Sagerer, “Person tracking with a mobile robot based on multi-modal anchoring,” in *Robot and Human Interactive Communication, 2002. Proceedings. 11th IEEE International Workshop on*, 2002, pp. 423 – 429.
- [51] R. Lienhart and J. Maydt, “An extended set of haar-like features for rapid object detection,” in *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 1, 2002, pp. I-900 – I-903 vol.1.
- [52] C. P. Papageorgiou, M. Oren, and T. Poggio, “A general framework for object detection,” in *Proceedings of the Sixth International Conference on Computer Vision*, ser. ICCV '98. Washington, DC, USA: IEEE Computer Society, 1998, pp. 555–. [Online]. Available: <http://portal.acm.org/citation.cfm?id=938978.939174>
- [53] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, june 2005, pp. 886 –893 vol. 1.

BIBLIOGRAPHY

- Bertozzi, M., Broggi, A., Grisleri, P., Graf, T., and Meinecke, M., “Pedestrian detection in infrared images,” in *Intelligent Vehicles Symposium, 2003. Proceedings. IEEE*, june 2003, pp. 662 – 667.
- Blackburn, M. R. and Nguyen, H. G., “Autonomous visual control of a mobile robot,” in *Proceedings of the 1994 Image Understanding Workshop, 1994*, pp. 1143–1150.
- Bradski, G. R., “Computer vision face tracking for use in a perceptual user interface,” 1998.
- Brethes, L., Menezes, P., Lerasle, F., and Hayet, J., “Face tracking and hand gesture recognition for human-robot interaction,” in *Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on*, vol. 2, 26-may 1, 2004, pp. 1901 – 1906 Vol.2.
- Chen, Z. and Lee, H.-J., “Knowledge-guided visual perception of 3-d human gait from a single image sequence,” *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 22, no. 2, pp. 336 –342, mar/apr 1992.
- Cheng, Y., “Mean shift, mode seeking, and clustering,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 17, no. 8, pp. 790 –799, Aug. 1995.
- Collins, R., “Mean-shift blob tracking through scale space,” in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 2, june 2003, pp. II – 234–40 vol.2.
- Comaniciu, D. and Ramesh, V., “Mean shift and optimal prediction for efficient object tracking,” in *Image Processing, 2000. Proceedings. 2000 International Conference on*, vol. 3, 2000, pp. 70 –73 vol.3.
- Comaniciu, D., Ramesh, V., and Meer, P., “Real-time tracking of non-rigid objects using mean shift,” in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, vol. 2, 2000, pp. 142 –149 vol.2.
- Comaniciu, D. and Meer, P., “Mean shift: A robust approach toward feature space analysis,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, pp. 603–619, May 2002. [Online]. Available: <http://portal.acm.org/citation.cfm?id=513073.513076>
- Comaniciu, D., Ramesh, V., and Meer, P., “Kernel-based object tracking,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, pp. 564–575, May 2003.

- Dalal, N. and Triggs, B., "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, june 2005, pp. 886–893 vol. 1.
- Doermann, D. and Mihalcik, D., "Tools and techniques for video performance evaluation," in *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, vol. 4, 2000, pp. 167–170 vol.4.
- Drarni, J. and Roy, S., "A simple oriented mean-shift algorithm for tracking," in *Image Analysis and Recognition*, ser. Lecture Notes in Computer Science, Kamel, M. and Campilho, A., Eds. Springer Berlin / Heidelberg, 2007, vol. 4633, pp. 558–568.
- Duraiswami, R. and Davis, L., "Efficient Mean-Shift Tracking via a New Similarity Measure," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pp. 176–183, 2005.
- Feyrer, S. and Zell, A., *Robust real-time pursuit of persons with a mobile robot using multisensor fusion*. I O S PRESS, 2000, pp. 710–715.
- Fox, D., Burgard, W., and Thrun, S., "Markov localization for mobile robots in dynamic environments," *Journal of Artificial Intelligence Research*, vol. 11, 1999.
- Fukunaga, K. and Hostetler, L., "The estimation of the gradient of a density function, with applications in pattern recognition," *Information Theory, IEEE Transactions on*, vol. 21, no. 1, pp. 32–40, Jan. 1975.
- Gavrila, D. M. and Davis, L. S., "Towards 3-d model-based tracking and recognition of human movement: a multi-view approach," *Int Workshop on Face and Gesture Recognition*, pp. 3–8, 1995. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.56.5329&rep=rep1&type=pdf>
- H. Zhang, L. L. and Yu., Q., "A scale rotation adaptive new mean shift tracking method," p. 68330S, 2008.
- Haritaoglu, I., Harwood, D., and Davis, L., "W4: Who? when? where? what? a real time system for detecting and tracking people," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, apr 1998, pp. 222–227.
- Huber, E. and Kortenkamp, D., "Using stereo vision to pursue moving agents with a mobile robot," in *Robotics and Automation, 1995. Proceedings., 1995 IEEE International Conference on*, vol. 3, may 1995, pp. 2340–2346 vol.3.
- Isard, M. and Blake, A., "Condensation – conditional density propagation for visual tracking," 1998.

- Kahn, R., Swain, M., Prokopowicz, P., and Firby, R., “Gesture recognition using the perseus architecture,” in *Computer Vision and Pattern Recognition, 1996. Proceedings CVPR '96, 1996 IEEE Computer Society Conference on*, jun 1996, pp. 734 –741.
- Kleinehagenbrock, M., Lang, S., Fritsch, J., Lomker, F., Fink, G., and Sagerer, G., “Person tracking with a mobile robot based on multi-modal anchoring,” in *Robot and Human Interactive Communication, 2002. Proceedings. 11th IEEE International Workshop on*, 2002, pp. 423 – 429.
- Kluge, B., Kohler, C., and Prassler, E., “Fast and robust tracking of multiple moving objects with a laser range finder,” in *Robotics and Automation, 2001. Proceedings 2001 ICRA. IEEE International Conference on*, vol. 2, 2001, pp. 1683 – 1688 vol.2.
- Kortenkamp, D., Huber, E., Bonasso, R. P., and Inc, M., “Recognizing and interpreting gestures on a mobile robot,” in *In Proceedings of AAAI-96*. AAAI Press/The MIT Press, 1996, pp. 915–921.
- Leung, A. P. and Gong, S., “Mean shift tracking with random sampling,” in *Proc. BMVC 2005*, 2006, pp. 729–738.
- Lienhart, R. and Maydt, J., “An extended set of haar-like features for rapid object detection,” in *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 1, 2002, pp. I–900 – I–903 vol.1.
- Lindeberg, T., “Feature detection with automatic scale selection,” *Int. J. Comput. Vision*, vol. 30, pp. 79–116, November 1998. [Online]. Available: <http://portal.acm.org/citation.cfm?id=305297.305298>
- Lipton, A., Fujiyoshi, H., and Patil, R., “Moving target classification and tracking from real-time video,” in *Applications of Computer Vision, 1998. WACV '98. Proceedings., Fourth IEEE Workshop on*, oct 1998, pp. 8 –14.
- Meis, U., Ritter, W., and Neumann, H., “Detection and classification of obstacles in night vision traffic scenes based on infrared imagery,” in *Intelligent Transportation Systems, 2003. Proceedings. 2003 IEEE*, vol. 2, oct. 2003, pp. 1140 – 1144 vol.2.
- Nanda, H. and Davis, L., “Probabilistic template based pedestrian detection in infrared videos,” in *Intelligent Vehicle Symposium, 2002. IEEE*, vol. 1, june 2002, pp. 15 – 20 vol.1.
- Niyogi, S. and Adelson, E., “Analyzing gait with spatiotemporal surfaces,” in *Motion of Non-Rigid and Articulated Objects, 1994., Proceedings of the 1994 IEEE Workshop on*, nov 1994, pp. 64 –69.

- Papageorgiou, C. P., Oren, M., and Poggio, T., “A general framework for object detection,” in *Proceedings of the Sixth International Conference on Computer Vision*, ser. ICCV '98. Washington, DC, USA: IEEE Computer Society, 1998, pp. 555–. [Online]. Available: <http://portal.acm.org/citation.cfm?id=938978.939174>
- Paragios, N. and Deriche, R., “Geodesic active regions for motion estimation and tracking,” in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 1, 1999, pp. 688–694 vol.1.
- Parzen, E., “On estimation of a probability density function and mode. the annals of mathematical statistics,” pp. 1065–1076, 1962.
- Peng, N. S., Yang, J., and Liu, Z., “Mean shift blob tracking with kernel histogram filtering and hypothesis testing,” *Pattern Recogn. Lett.*, vol. 26, pp. 605–614, April 2005. [Online]. Available: <http://dx.doi.org/10.1016/j.patrec.2004.08.023>
- Rohr, K., “Towards model-based recognition of human movements in image sequences,” *CVGIP: Image Understanding*, vol. 59, no. 1, pp. 94 – 115, 1994. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1049966084710060>
- Rohr, K., “Towards model-based recognition of human movements in image sequences,” *CVGIP: Image Underst.*, vol. 59, pp. 94–115, January 1994. [Online]. Available: <http://portal.acm.org/citation.cfm?id=183817.183826>
- Schlegel, C., Illmann, J., Jaberg, H., Schuster, M., and Worz, R., “Vision based person tracking with a mobile robot,” in *In Proc. British Machine Vision Conf*, 1998, pp. 418–427.
- Schulz, D., Burgard, W., Fox, D., and Cremers, A., “Tracking multiple moving objects with a mobile robot,” in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, 2001, pp. I-371 – I-377 vol.1.
- Sidenbladh, H., “Probabilistic tracking and reconstruction of 3D human motion in monocular video sequences,” Numerical Analysis and Computer Science, Doctoral Dissertation, ISRN KTH/NA/P-01/14-SE, Oct. 2001.
- Smith, K., Gatica-Perez, D., and Odobez, J.-M., “Using particles to track varying numbers of interacting people,” in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, june 2005, pp. 962 – 969 vol. 1.
- Smith, K., Gatica-perez, D., marc Odobez, J., and Ba, S., “Evaluating multi-object tracking,” in *In Workshop on Empirical Evaluation Methods in Computer Vision*, 2005.

- Sun, S., Haynor, D., and Kim, Y., “Semiautomatic video object segmentation using vsnakes,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 1, pp. 75 – 82, Jan. 2003.
- Turk, M. and Pentland, A., “Face recognition using eigenfaces,” in *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR '91., IEEE Computer Society Conference on*, jun 1991, pp. 586 –591.
- Viola, P. and Jones, M., “Rapid object detection using a boosted cascade of simple features,” in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, 2001, pp. I-511 – I-518 vol.1.
- Waldherr, S., Thrun, S., Romero, R., and Margaritis, D., “Template-based recognition of pose and motion gestures on a mobile robot,” in *In Proceedings of the AAAI Fifteenth National Conference on Artificial Intelligence.* MIT Press, 1998, pp. 977–982.
- Wren, C., Azarbayejani, A., Darrell, T., and Pentland, A., “Pfinder: real-time tracking of the human body,” in *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, oct 1996, pp. 51 –56.
- Xu, F., Liu, X., and Fujimura, K., “Pedestrian detection and tracking with night vision,” *Intelligent Transportation Systems, IEEE Transactions on*, vol. 6, no. 1, pp. 63 – 71, march 2005.
- Yilmaz, A., Li, X., and Shah, M., “Contour-based object tracking with occlusion handling in video acquired using mobile cameras,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 11, pp. 1531 –1536, nov 2004.
- Yilmaz, A., Javed, O., and Shah, M., “Object tracking: A survey,” *ACM Comput. Surv.*, vol. 38, no. 4, p. 13, 2006.
- Zhao, L. and Thorpe, C., “Stereo- and neural network-based pedestrian detection,” in *Intelligent Transportation Systems, 1999. Proceedings. 1999 IEEE/IEEJ/JSAI International Conference on*, 1999, pp. 298 –303.