OBJECT TRACKING FOR SURVEILLANCE APPLICATIONS USING THERMAL AND VISIBLE BAND VIDEO DATA FUSION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

ÇİĞDEM BEYAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

DECEMBER 2010

Approval of the Graduate School of Informatics

_____

Prof. Dr. Nazife BAYKAL

Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

_____

Prof. Dr. Yasemin YARDIMCI

Head of Department

This is to certify that I have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

_____

Assist. Prof. Dr. Alptekin TEMİZEL

Supervisor

Examining Committee Members

Prof. Dr. Yasemin YARDIMCI (METU, II) _____

Assist. Prof. Dr. Alptekin TEMİZEL (METU, II) _____

Prof. Dr. Gözde BOZDAĞI AKAR (METU, EEE)_ _____

Assist. Prof. Dr. P. Erhan EREN (METU, II) _____

Assoc. Prof. Dr. Ziya TELATAR (Ankara Univ, EE) _____

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

Name, Last name: : Çiğdem Beyan

Signature : _____

ABSTRACT

OBJECT TRACKING FOR SURVEILLANCE APPLICATIONS USING THERMAL AND VISIBLE BAND VIDEO DATA FUSION

BEYAN, Çiğdem
M.S., Department of Information Systems
Supervisor: Assist. Prof. Dr. Alptekin Temizel

December 2010, 92 pages

Individual tracking of objects in the video such as people and the luggages they carry is important for surveillance applications as it would enable deduction of higher level information and timely detection of potential threats. However, this is a challenging problem and many studies in the literature track people and the belongings as a single object. In this thesis, we propose using thermal band video data in addition to the visible band video data for tracking people and their belongings separately for indoor applications using their heat signatures. For object tracking step, an adaptive, fully automatic multi object tracking system based on mean-shift tracking method is proposed. Trackers are refreshed using foreground information to overcome possible problems which may occur due to the changes in object's size, shape and to handle occlusion, split and to detect newly emerging objects as well as objects that leave the scene. By using the trajectories of objects, owners of the objects are found and abandoned objects are detected to generate an alarm. Better tracking performance is

also achieved compared a single modality as the thermal reflection and halo effect which adversely affect tracking are eliminated by the complementing visible band data.

Keywords: Mean Shift, Abandoned Object Detection, Automatic Multiple Object Individual Tracking, Visible Band Image, Thermal Image

ÖZ

TERMAL VE GÖRÜNÜR VİDEO BAND VERİ TÜMLEŞTİRME KULLANARAK GÖZETLEME UYGULAMALARI İÇİN NESNE TAKİBİ

BEYAN, Çiğdem
Yüksek Lisans, Bilişim Sistemleri Bölümü
Tez Yöneticisi: Yrd. Doç. Dr. Alptekin Temizel

Aralık 2010, 92 sayfa

Gözetleme uygulamalarında, insanlar ve taşıdıkları bagajlar gibi nesnelerin videoda ayrı takibi nesnelerin gezingeleri ve birbirleriyle olan etkileşimi gibi yüksek seviye bilgilerin çıkarılması ve olası tehditlerin zamanında tespitine olanak kılması nedeniyle önemli bir yer tutar. Ancak, nesnelerin ayrı takibi zor bir problemdir ve literatürdeki birçok çalışma insanların ve onların eşyalarının tek bir nesne gibi takibini önermektedir. Bu tezde, görünür bant video verisine ek olarak termal bant video verisi kullanılmış ve nesnelerin sınıflandırılması için sıcaklık bilgisinden yararlanılmıştır. Nesnelerin ayrı takibi için nesneleri insan ve taşıdıkları nesneler şeklinde ayıran bir iç alan uygulaması sunulmuştur. Nesne takibi için, ortalama değer kayması takibi tabanlı, uyarlanabilir, tamamen otomatik, çoklu nesne takip sistemi önerilmiştir. Takipçiler, nesnenin boyutunun, şeklinin değişmesi, nesnenin bir başka

nesne tarafından önünün kapanması, nesnelerin ayrılması ve sahneden ayrılan nesneler kadar sahneye yeni giren nesnelerin de saptanması gibi durumlarda olabilecek problemleri aşmak için ön plan bilgisi kullanılarak yenilenir. Nesnelerin gezingeleri kullanılarak, eşyaların sahipleri bulunmuş ve terk edilen nesneler tespit edilerek alarm verilmiştir. Tek tarz veri kullanarak gerçekleştirilen uygulamalarla karşılaştırıldığında nesne takibini olumsuz etkileyen termal yansıma ve "halo" etkisi gibi problemler görülebilir bant veriyi tamamlayıcı olarak kullanarak ortadan kaldırılmış ve daha iyi bir takip performansı elde edilmiştir.

Anahtar Kelimeler: Ortalama Değer Kayması, Terk Edilmiş Nesne Tespiti, Otomatik Uyarlanabilir Çoklu Nesne Ayrık Takibi, Görünür Bant İmge, Termal İmge

ACKNOWLEDGMENTS

I would like to state my special thanks and gratitude to Assist. Prof.Dr. Alptekin Temizel for his supervision, encouragement, patience and support at all phases of my thesis. Under his guidance, I have chance to develop myself both academically and personally.

I would like to thank to Prof. Dr. Yasemin Yardımcı, Prof. Dr. Gözde Bozdağı Akar, Assist.Prof.Dr. Erhan Eren and Assoc.Prof.Dr. Ziya Telatar for accepting to read and review this thesis and for their valuable contributions.

I am also thankful to my friends, Ahmet Yiğit, Fatih Ömrüuzun, Püren Güler, Deniz Emeksiz, Mustafa Teke and Ersin Karaman for their help in capturing the test videos.

I would like to thank to my friends (Mehmet Dikmen, Oya Çınar, Tülin Erçelebi, Öykü Eren, Burçak Sönmez, Seda Şahin, Didem Tokmak, Serian Yazgı…) for their endless support and motivation throughout the thesis.

Finally, my biggest gratitude is to my family (Esin Beyan, Cengiz Beyan, Süreyya Ünlü and Sadık Ünlü) for their endless patience, support, love and trust. I am proud to be your daughter and grandchild.

To My Family…

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

1D	: One Dimensional
3D	: Three Dimensional
CCTV	: Closed Circuit Television
CM	: Combination Model
CSM	: Contour Saliency Map
DV	: Digital Video
FN	: False Negative
FOV	: Field of View
FP	: False Positive
GFM	: General Fusion Model
GPU	: Graphics Processing Unit
HOG	: Histogram Oriented Gradient
HSV	: Hue, Saturation, Value
I	: Infrared Band
JPEG	: Joint Photographic Experts Group
LIO	: Local Intensity Operation
LWIR	: Low Wavelength Cameras

MAT	: Mean Absolute Thresholding
MWIR	: Medium Wavelength Infrared Cameras
NIR	: Near Infrared Cameras
OpenCV	: Open Computer Vision
PAL	: Phase Alternate Line
PCA	: Principal Component Analysis
pdf	: Probability Density Function
RGB	: Red, Green, Blue
ROI	: Region of Interest
SVM	: Support Vector Machine
TP	: True Positive
VSAM	: Video Surveillance and Monitoring
YCrCb	: Luminance, Chrominance

CHAPTER 1

INTRODUCTION

Individual tracking of objects such as people and the belongings they carry is an important task for video surveillance applications as it would allow making higher level inferences and timely detection of threats. However, although this task is important, it is also a challenging problem and studies in literature track people and belongings as a single object. Object tracking on the other hand, is an important and challenging task in many computer vision applications as it is the base of individual tracking of objects. In the past few decades, various object tracking algorithms have been proposed.

Mean-shift tracking plays an important role in this area due to its robustness, ease of implementation and computational efficiency. However, the standard mean shift algorithm has a number of shortcomings which affect tracking performance and cause inaccurate or false tracking. In the literature, mean shift tracking based methods generally focus on a single drawback of this object tracking method. However, to achieve a robust tracking, all the drawbacks need to be addressed. Additionally, most of these methods require manual object detection to determine the objects that will be tracked. Besides, overwhelming majority of studies aim to track only one object at a time and do not propose a solution for tracking of multiple objects.

In this thesis, a fully automatic multiple object tracker based on mean shift algorithm is presented. Foreground detection is used to initialize object trackers. The bounding box of the object, obtained as the result of background subtraction and object segmentation steps, is used as a mask to make the system more efficient by

decreasing the number of iterations to converge for the new location of the object. Additionally, using background detection, new objects entering the field of view (FOV) and objects that are leaving the scene could be detected. Trackers are refreshed to solve the potential problems which may occur due to changes in objects' size, shape, to handle occlusion, split and to detect newly emerging objects as well as objects that leave the scene.

This proposed object tracking method is applied on only visible band. A shadow removal method is used to increase the tracking accuracy of the method in visible domain. On the light of the results of this method applied in visible domain, the proposed method remedies problems of mean shift tracking and presents an easy to implement, robust and efficient tracking method that can be used for automated video surveillance applications. Furthermore, it is shown that the proposed method is superior to the standard mean shift. In addition to the application in only visible domain, the proposed method is also applied in thermal band only. Similar results are obtained in this modality as well.

In this thesis, in addition to the object tracking method based on mean shift tracking, we also propose using fusion of thermal imagery to the visible band imagery for tracking in indoor applications such as airports, metro and railway stations. We use the approach mentioned above to track multiple objects with a fully automatic mean shift tracking algorithm and visible, thermal domain tracking information are fused to allow tracking of people and the object they carry separately using their heat signatures. By utilizing the trajectories of these objects, interactions between them could be found and threats such as abandoning of an object by a person could be detected. Better tracking performance when compared to using single modality is also achieved as drawbacks of using single modality are eliminated. The proposed method has been tested on videos containing various scenarios. The experimental results show that the presented method is effective for separate tracking of objects and for detecting the interactions in the presence of occlusions.

Detection of packages that are left unattended in public spaces such as shopping malls, railways and airports is related to security and important to prevent possible hazards. To control the security of environment, surveillance operators generally

watch a high number of cameras simultaneously. However, this is a challenging and labor intensive task. Additionally, these systems are left unattended at certain times which results in security vulnerability and may cause tragic events. Therefore, it is crucial to have automated abandoned object detection systems aiding operators in place to detect unattended suspicious items on time with low false alarm rates.

In this thesis, we propose an abandoned object detection method based on the multimodal approach for individual tracking of people and their belongings explained above. Utilizing the trajectories of the people and their belongings, information coming from individual tracking, owners of the belongings is found and abandoned objects are detected to generate an alarm.

The remaining part of the thesis is organized in six parts as follows. Chapter 2 presents an overview of related works in the literature which contain algorithms that could be utilized to track objects. Our methods for object tracking called adaptive mean shift for automated multi object tracking is explained and application of this algorithm in only visible band and only thermal band are presented in Chapter 3. Our approach for individual tracking of people and their belonging and abandoned object detection (based on the object tracking method presented in Chapter 3) using fusion of thermal and visible modalities is explained in Chapter 4. The strengths, the weaknesses and comparisons of these methods (methods proposed in Chapter 3 and 4) with the existing studies and experimental results in different scenarios are given in Chapter 5. Finally, in Chapter 6 the thesis is summarized and it is concluded by suggestions for future works.

CHAPTER 2

AN OVERVIEW OF RELATED WORKS

Object tracking using thermal and visible band data fusion requires some methods such as background subtraction, shadow removal, data fusion and so forth.

In this chapter, the related works which includes algorithms that could be utilized to track objects are summarized. This chapter is divided into six parts as background subtraction, shadow detection, data fusion, object tracking, individual tracking of objects and abandoned object detection.

2.1. Background Subtraction

Background subtraction is a technique for segmenting foreground objects and it is widely used in video surveillance applications. It helps to detect changes in the environment and to track the moving objects in the environment. It aims to detect stationary objects by subtracting the current image from a reference background image. Although many background subtraction methods are successful to detect foreground image and provide robust systems, these methods sometimes suffer from problems and challenges while detecting background. Changes in illumination, shadows and noise which may result in camera movement are the most important and frequently encountered problems. Additionally, moving objects and weather conditions (rain, snow…) causing non-static background are other challenges. In order to provide a robust system by segmenting the objects correctly, background subtraction algorithms must be adaptive to changes and successful in handling such challenges.

Temporal differencing is a basic method and tries to detect moving pixels by applying pixel-by-pixel differencing to adjacent frames in a video sequence. This method is frequently used due to its ease of implementation and it is successful and adaptable to detect dynamic scene changes. However, it generally fails to detect the entire pixels of objects and some holes inside of the object (which may cause false object detection in some applications) may occur. Additionally, it is not successful in detecting stationary objects in the scene, generally classifies these objects as background and also fails if these stationary objects start to move. In study [1], a two frame differencing scheme was presented. According to this method only the current frame and the previous frame are utilized. Each pixel's intensity is subtracted from the corresponding previous frame pixel's intensity and if this difference is greater than a threshold, then this pixel is classified as foreground pixel otherwise it is assigned as background pixel. Since, this method cannot cope with multi modal background distributions and cannot handle the challenges which are discussed above, other variations such as finding the average or median of the previous N frames to estimate background were proposed to overcome these drawbacks. For instance, study [2] proposed a three frame differencing approach instead of two frame differencing by combining this technique with an adaptive background subtraction model for Video Surveillance and Monitoring (VSAM) project. In this method, a pixel is classified as moving pixel (foreground pixel), if the difference between the intensities of that pixel in current and previous frames and the difference between the intensities of that pixel in current and next-to-previous frame are higher than a threshold, otherwise pixel is determined as background pixel. This algorithm successfully segments moving objects when compared to two frame differencing method [1], however it cannot detect and segment the interior pixels of a moving object. On the other hand, a method called temporal median filter which is based on median of last N frames was discussed in [3]. Although this method is easy to implement, it has the disadvantage of requiring a buffer in order to store pixel values to find the median.

An eigenvalue decomposition method is applied to detect background in [4]. In this study, eigenvalue decomposition technique was applied to the whole image as opposed to the studies [1-3] which apply their algorithms pixel-by-pixel. In this

method, mean image is calculated by using *N* frames in the training step. Then covariance matrix is computed using this mean image and best eigenvectors are obtained and stored as an eigenvector matrix. Then, a classification step is applied, the test image is projected onto the eigenvectors by using eigenvector matrix and projected image is constructed. Finally, the difference between the new test image and projected image is calculated and the foreground pixels are determined as the positions that the difference is greater than a threshold. According to the experimental results in [3], this method [4] is superior to studies [1] and eigenvalue decomposition method given in [3] in terms of segmentation accuracy, its memory load depends on number of frames which are used in training phase and the speed changes depending on the number of best eigenvectors.

Even though the methods that are mentioned above are often successful and have an acceptable accuracy in specific applications, they generally do not overcome the problems and challenges such as non-static background, different weather conditions, camera noise and so forth which affect the robustness of background subtraction. Additionally, these algorithms cannot handle multimodal background distributions. Therefore, in many studies such as [5], Mixture of Gaussians technique has been utilized. This technique is one of the most powerful background subtraction methods if background changes very fast and more than one background distribution exist. Furthermore, it is reliable while computationally not very complex. This method is a pixel-based background subtraction method and the probability of a pixel value *x* at time *t* is defined by mixture of *K* Gaussian distributions as follows:

$$P(x_t) = \sum_{i=1}^{K} \omega_{i,t}\, \alpha I_t + \eta(x_t, \mu_{i,t}, \sum_{i,t})$$
$$\eta(x_t, \mu_t, \sum_t) = \frac{1}{(2\pi)^{n/2} |\sum|^{1/2}} e^{\frac{1}{2}(x_t - \mu_t)^T \sum^{-1} (x_t - \mu_t)} \tag{2.1}$$

where $\omega_{i,t}$ is weight, μ_t is the mean at time *t*, $\sum_{i,t}$ is the covariance matrix of the i^{th} Gaussian at time t, η is the Gaussian probability density function, and α is learning parameter.

Firstly, all the distributions are ranked according to ratio between the peak amplitude ω_i and standard deviation σ_i. Then, the first B distributions in the ranked which are satisfying Eq. 2.2 is defined as background. While defining background, the higher and more compact distribution is defined as likely to belong to the background.

$$\sum_{i=1}^{B} \omega_i > T \tag{2.2}$$

This method could find the background and foreground pixels without requiring large number of video frames. It is successful in overcoming lighting changes, repetitive motions such as tree leaves moving in the wind, noise, multiple surfaces for a particular pixel at the same time and finding slowly moving objects. The foreground objects could completely be extracted (both internal and bounding pixels of the object) and its accuracy is higher than the methods that are mentioned above [3].

An Improved Adaptive Gaussian Mixture model which is based on mixture of Gaussians is proposed in [6]. This algorithm has been reported to be a reliable background subtraction method while computationally not very complex [6]. This method is also a pixel-based background subtraction technique like mixture of Gaussians method and a pixel could belong either to background or foreground. This method is powerful to detect background and foreground pixels, cope with the problems and challenges that are mentioned above and could completely extract the foreground objects. Moreover, since this method does not use a fixed number of components, it is proposed as more adaptive and robust when compared to mixture of Gaussians method. In other words, Improved, Adaptive Gaussian Mixture model [6] could automatically select the proper number of components per pixel and updates parameters (mean μ, standard deviation σ…). In this thesis we decided to use this technique because of the explained advantages. The algorithm of the method is defined in more detail in Section 3.1.2.

2.2. Shadow Detection

Most of the background subtraction and foreground detection algorithms are susceptible to shadows which cause the object detection methods fail in object segmentation. Shadows might cause serious problems such as merging of objects, shape deformations, distortions of color histogram of objects, and make the next steps such as object classification and object tracking to perform inaccurately. In the literature, generally two kinds of shadows are defined. Cast shadow is defined as the area projected by the object in the direction of light while self shadow is defined as a part of the object which is not illuminated by direct light [7].

The methods in literature generally use chromaticity, brightness values and prefer to use Hue, Saturation, Value (HSV) or Red, Green, Blue (RGB) color spaces to handle shadows. In [8], each pixel is represented by a color model which separates brightness from the chromaticity component. A pixel is classified into four classes: background, shadow, highlighted background and moving object by calculating the change of brightness and chromaticity between the background and the current image pixels. In this study, it is assumed that the shadows pixel's RGB color vectors are in the same direction with RGB color vectors of the corresponding background pixel while the shadow pixel's brightness value is less than the corresponding background pixel's brightness value. In another study [9]; similar approach is applied, chromaticity and intensity gradient information are used. To determine the pixels belonging to shadow the change in intensity and chromaticity are calculated. If there is a change in intensity while there is not much change in chromaticity, that pixel is classified as shadow. In addition to that, the gradient information is utilized to moving pixels in order to provide much more reliable technique. Another shadow detection technique for tracking people is illustrated in [10]. Different from the other techniques which utilize the brightness and chromaticity and classify the shadow according to decrease in these values, this method only uses mean values of each band instead of multiple samples of the bands. To determine the shadow, the change in luminance and chrominance are calculated and if these changes are between predefined thresholds then the pixel is defined as shadow pixel. In [11], HSV color space is used. As a first step, the shadow pixels are detected using the assumption that the shadow decreases the luminance and saturation values while does not

significantly change the hue component. Then, local neighborhoods of these shadow pixels are checked; if the illumination ratio of two shadow pixels is not similar then these pixels are assigned as unclassified. The number of foreground, shadow and unclassified pixels are calculated in the next step and with the help of some heuristic, unclassified pixels are also classified as foreground or shadow.

In this thesis, we used a shadow detection scheme which is based on the HSV color space and presented in [12]. HSV color space is selected as in [12] this color space is reported as a more robust color space than RGB since it separates chromaticity and luminosity and it has been shown that the mathematical formulation of shadow detection is easier than in RGB space. While finding the shadow pixels, the fact that shadow cast on a background does not significantly change its hue and saturation information while shadows decrease the saturation of the pixels is used. The details of the algorithm are given in Section 3.1.3.

2.3. Data Fusion

Tracking systems typically utilize single modality of video in the visible band. These systems are successful in controlled conditions but their performance degrade when there are instant lighting changes, shadows, smoke and unstable backgrounds. Additionally, they might not work well when the foreground object has similar coloring to the background. These conditions bring false positive detections and also false tracking. On the other hand, with lowering cost of thermal cameras, which has predominantly been used in military applications due to their costs until recently, thermal imagery have become more feasible and started to be utilized for civil applications. Thermal cameras, on the contrary to visible band cameras, are not affected by lighting changes, illumination, shadows, darkness and so forth. Thermal cameras generate images depending on the amount of emitted thermal radiation by the objects. The emitted radiation and the temperature of the object are related and emitted radiation increases when the temperatures of the objects such as humans and/or animals go up. Therefore, these hot objects appear much brighter when compared with cooler objects against a cooler environment (assuming white-hot setting). On the other hand, since these cameras use infrared energy for imaging like near infrared cameras (NIR), they are sometimes confused with NIR cameras.

However, NIR cameras measure the infrared light reflected by objects and work at 0.7-1 μm infrared spectrum. Similarly, visible band cameras also measure the reflected light but work in 0.38-0.75 μm spectrum and both of them require illumination to capture images. Thermal cameras, on the contrary, are passive and do not require any kind of illumination. Low wavelength infrared cameras (LWIR) on the other hand, measure the thermal radiation emitted by the objects typically at 8-14 μm range and medium wavelength infrared cameras (MWIR) which uses more expensive cooled sensors capture 3-8 μm range. However, they sometimes cannot detect objects when objects' thermal properties are similar to the environment's thermal properties. "Halo effect" which appears around very hot or dark objects is another disadvantage of thermal video [13]. Additionally, thermal reflection is a different source of problem. Surfaces such as wet, glass, metals reflect infrared radiation and may cause false alarms while detecting or tracking objects when just thermal technology is used. While visible and thermal band have their own limitations, using modalities from both visible and thermal band in other words fusion of these two kinds of data in surveillance systems allow us to overcome the drawbacks of these two kinds of modalities and obtain more robust systems.

In the literature, fusion is applied in different stages of object tracking. For instance some studies fuse images during motion detection, others fuse the motion detection results of thermal and visible images, or fuse the results of two object trackers coming from thermal and visible band [14]. In this thesis, we applied fusion after object detection for object tracking as in a recent study, it has been reported that fusion after object detection approach is the most successful scheme [14]. On the other hand, studies which utilize fusion of thermal and visible imagery for different domains also exist. For instance, in [15] a region-based algorithm that finds and fuses most salient contours coming from thermal and visible modality to present a better background subtraction is proposed. In this study, contours in a selected region of interest (ROI) from either domain are extracted and combined into a fused contour image. Fused contours are then post processed to obtain silhouettes. Contour Saliency Map (CSM) which shows the degree of that pixel belonging to the boundary of object is constructed. CSM for all ROIs in thermal and visible domains are computed and fused applying union operation. On the other hand, a surveillance

system that fuses thermal infrared and visible spectrum video for pedestrian detection and tracking is presented in [16]. In this study, fusion is performed at the object level. Different from the many studies which aim pedestrian tracking, in this study detection and tracking are applied separately in two kinds of modality by using result of background subtraction, a rule based connected component linking and some heuristics. This work is successful when compared to traditional visible only spectrum studies but utilization of many heuristics may cause problems. Another approach that uses Fuzzy logic and Kalman filtering in the fusion step in order to detect moving objects is presented in [17]. In addition to presenting a different fusion example, this study introduced the measurement parameters to determine measurement accuracy of each sensor. A method that uses closed circuit television (CCTV) and thermal image fusion for object segmentation and tracking is focused on in [18] and [19]. In this method, background subtraction is applied separately in each modality and thresholding based fusion is done to obtain foreground objects for tracking. To fuse these two kinds of modality and segment foreground objects Transferable Belief Model is applied. Then extracted foreground objects are tracked by representing each pixel of the object with a multi dimensional Gaussian. Furthermore in [19], different fusion methods are investigated. According to this study, fusion methods can be divided into two groups as General Fusion Model (GFM) and Combination Model (CM). Fusion methods in GFM are based on pixel fusion and each pixel is modeled as mixture of Gaussians according to number of data sources. On the other hand, CM fusion methods is the model based which contains weighted averaging of models, similarity score product of models, minimum and maximum score fusion of models. In [20], a framework for detecting people using thermal and visible image sequences is proposed. In this framework, firstly, images coming from two thermal and two visible band camera are registered. Subsequently, object annotation is applied and bounding boxes of objects are extracted from registered images. Histogram of Oriented Gradient (HOG) features coming from visible and thermal domains, disparity based detectors which are constructed according to person size and depth and Support Vector Machine (SVM) are used for object classification and fusion of two modalities.

2.4. Object Tracking

Object tracking is an important and challenging task in many computer vision applications such as automated surveillance, motion-based recognition, human computer interaction, vehicle navigation, traffic monitoring, and autonomous robot navigation. The objective of tracking is to find the location of a given object or a part of the object in the current frame and establishing correspondence of objects and object parts between adjacent frames of video sequences. In other words, object tracking is estimating the trajectory of an object in the video sequences [21, 22]. It is a significant problem since it provides higher level data and could be use as a base for some applications such as activity analysis, behavior recognition and finding interactions between objects. Tracking objects can be difficult due to problems in object segmentation caused by noise in images, shadows, illumination changes. Furthermore, partial or full object occlusion, complex object motions, complex object shapes may cause object tracking to fail [21, 22].

Up to now, numerous object tracking methods have been proposed using different methodologies, different object representations, and features. In study [21], these different object tracking methods are summarized; the object representations and the features for single modality tracking that are frequently used by studies are given. According to this study, to represent objects; points (which are used as a set of points of object or the centroid of an object), geometric shapes (that used to represent objects by a rectangle, ellipse…), object contours (which define the boundary of an object), object silhouettes (that means the inside of the object) and articulated shape models (which contains body parts such as hands, legs, torso) are being used. On the other hand, probability density of object which could be Gaussian, mixture of Gaussians, Parzen window, histograms and so forth, templates using geometric shapes, active appearance models which contain an appearance vector in the form of color, texture or gradient magnitude, multi-view appearance models that contain different views of an object for instance using Principal Component Analysis (PCA) to represent shape of object are utilized to represent appearance features of objects [21]. Common features used by many studies in literature is also compared and analyzed in [21]. The color of the object is one of the most preferred features. As a color space RGB, YCrCb (luminance, chrominance), HSV could be used. However

according to [21] using RGB color space is not suitable since it is not a uniform color space and RGB components are highly correlated with each other, while HSV is an approximately uniform color space although it is sensitive to noise. In this thesis, we used YCrCb color space in the object tracking step which is defined as the most appropriate color space in [21]. Besides color feature, edges which are used to identify the boundary of object are generally used by methods tracking the boundary of the object. Optical flow defined as *"dense field of displacement vectors which defines the translation of each pixel in a region"* in study [21] is generally used as a feature for motion-based tracking. Texture which represents the spatial properties of the objects are favorable features and used by many studies.

In study [21] object tracking is divided mainly into three groups as it is shown in Figure 2.1.

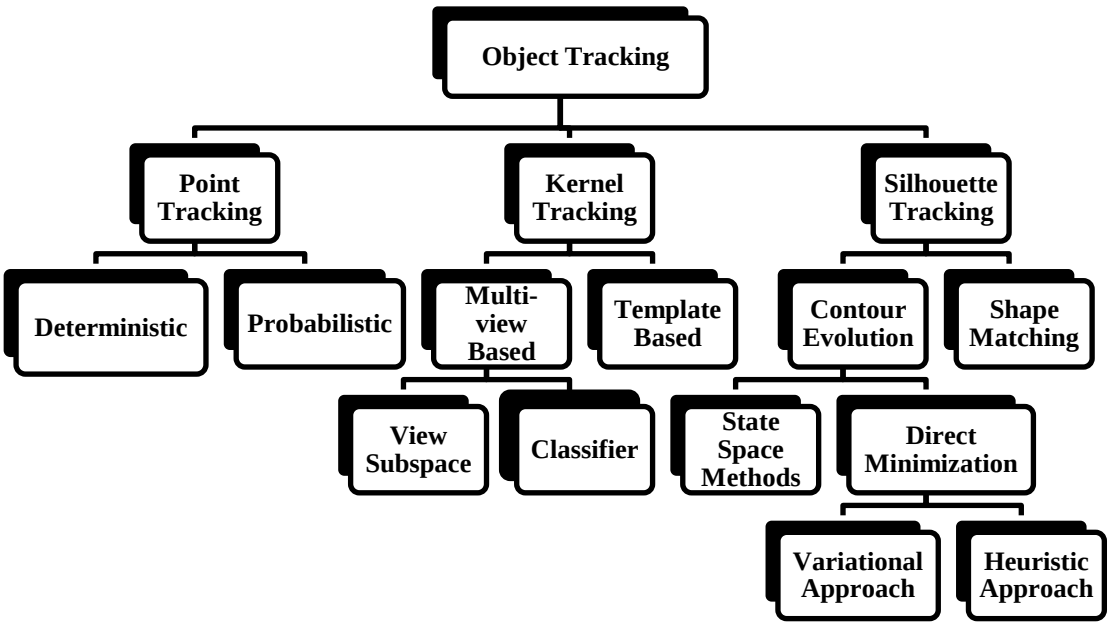


Figure 2.1: Categorization of object tracking methods [21]

In point tracking method, objects are represented by points and the matching of the points in the current frame is based on the object position and motion in the previous frame. Correspondence based tracking is a kind of point tracking. Kernel tracking on the other hand, is based on object shape and appearance. A kernel which could be a rectangular or elliptical shape is defined and the features such as color histogram of

the points inside of the kernel are used to match the tracked object and motion of the kernel is calculated to find the new position of the kernel in the consecutive frames. Mean-shift tracking is an example to kernel tracking. Differently, silhouette tracking is performed by finding the object region in each frame. This kind of tracking methods uses the pixels inside the object region. Edge information is generally used as a feature and the silhouettes are tracked by contour or shape matching.

As an example of point tracking, [22] proposes a correspondence based object tracking algorithm which does not track object parts such as limbs, torso but track objects as a whole frame by frame. In this study, size, center of mass, bounding box and color histogram are used as features to match the object in the adjacent frames. An occlusion and split algorithm is also presented which is based on the estimation of intersection of the bounding boxes of the objects. Although this study performs a successful and robust tracking, applying this method in every frame is not efficient since it has high memory requirement to hold the objects and also objects' correspondence objects. In [23], correspondence matching is combined with a motion estimation algorithm to track objects or object parts such as heads, hands, torso and feet. Cardboard Model [24] representing relative positions and sizes of object parts is used to track these body parts. To handle merge and split, appearance of objects are stored in order to match after splitting of objects. Parts of the body is tracked by using a model based tracking method in [25] which utilize color and shape features to track head and hands of human in real-time. An object segmentation using background subtraction and appearance model with a Particle Filter to track multiple objects is given in [27]. The relationship between the objects is extracted by matching the object blobs in the frames and trajectory of multiple objects are found in [28]. Firstly, connected components are extracted from foreground image and these connected components are stored in order to use in presence of merging or splitting. An inference graph is build to extract and store relations between the connected components. Using this graph and object model with object's spatial information the objects are classified as a whole object, a part of an object or a group of objects.

Mean shift algorithm [28] is a popular technique due to its robustness, ease of implementation and computational efficiency. However, the standard mean shift

algorithm suffers from a number of problems which adversely affect tracking performance and could cause inaccurate or even false tracking. Firstly, it is not adaptable to changes in objects' size or shape and its performance is dependent on correct kernel size selection in the object initialization phase. Another shortcoming is the inclusion of background information into the object model as the kernel shape does not always fit the object. Additionally, the tracking might shift and even fail when object is occluded or background colors are similar to the foreground objects' colors. Many studies have attempted to improve on mean shift to solve these problems. For instance, additional spatial information to mean shift tracking was used in [29] to obtain a better description of target object in order to increase the robustness of tracking. In this study, usage of spatiogram, which is a kind of histogram, where each histogram bin is weighted by the spatial mean and covariance of pixels, was presented. The inclusion of spatiogram to standard mean shift was reported to provide a better matching performance. In [30], a method is presented to solve the problems which may occur when background color is similar to the color of the object that will be tracked using the background position on the previous frame and current frame to compute the target model. Although this method has low computational load, it does not handle occlusions, merging of objects and tracking of multiple objects. A study to solve the problems which may arise due to incorrect mean-shift kernel scale selection is addressed in [31]. This study used difference of Gaussians kernel and provides a good tracking performance by handling changes in target scale. However, it requires high computational cost and is not suitable for real time applications. To adapt the kernel scale and the orientation of kernel, various approaches have been proposed. For instance, [32] combines the mean shift method with adaptive filtering. Even though the kernel scale and orientation estimations are successful due to the use of symmetric kernels, actual object shape might not be matched. An alternative human tracking method which uses multiple radially symmetric kernels is proposed in [33]. In this study, a flexible tracking method was presented which allows optimization of kernel parameters for a specific class of objects. This study is useful especially when larger kernel sizes such as arms, legs and/or smaller kernel sizes such as torso are needed to be tracked. On the other hand, [34] introduces an adaptive asymmetric kernel which is able to deal with out of plane

rotations by using some heuristic. Another adaptive mean shift tracking using multi-scale images is presented in [35]. In this study, Gaussian kernel is preferred and kernel bandwidth is determined by using a log-likelihood function. Although this method estimates the exactly position of the tracked object, it would not be efficient if it is used in real time applications and for multiple object tracking due to the fact that it needs nearly three iterations per object to converge to the correct object position. A method for multiple objects tracking and try to solve the problem of inclusion of background information into the object model which may results in when the relocation of an object is large is given in [36]. Multiple kernels were utilized in moving areas, background and template similarities were used to improve the convergence of tracker.

In the literature, mean shift tracking based methods generally focus on a single shortcoming of mean shift. However, to achieve a robust automated tracking, all the problems need to be handled. Most of these methods require manual object identification and need human input to define the object that will be tracked. Besides, overwhelming majority of studies aim to track only one object at a time and do not provide a solution for tracking of multiple objects. In this thesis, we propose an easy to implement and fully automatic multiple object tracking algorithm based on standard mean shift method (Chapter 3).

2.5. Individual Tracking of Objects

Individual tracking of objects such as people and the luggage they carry is an important issue for video surveillance applications as it would allow making higher level inferences and make timely detection of potential threats, such as abandoning of an object by a person. Additionally, by using the trajectories of these living and nonliving objects, interactions between them could be deduced. However, this is a challenging problem and in the literature, people and objects they carry are tracked as a single object. An abandoned object detection method and abandoned object owner matching are proposed in [14, 37]. However, neither of these studies could track people and their belongings separately while the object is being carried. The method in [37] finds the owner of object when the object and its owner split which

results in false owner association and cause false alarm in the case of occlusion of owner or abandoned object. Besides these, two people who are entering the scene together may cause a false alarm when they split since one of them will be detected as an abandoned object. Additionally, for this case, it would be difficult to discriminate the object by using the size of object since the luggage and for instance a child's sizes could be very similar. Differently, [14] firstly detects the abandoned object, then searches through the history to find the owner of the abandoned object and associate the people with abandoned object by calculating the overlap of the bounding boxes of owner and abandoned objects. But this method is not efficient and requires too much memory since it stores locations of each tracked object for all frames.

2.6. Abandoned Object Detection

In recent years, several studies have been proposed to detect abandoned objects automatically by using computer assisted surveillance systems. In such systems, the aim is detecting all real alarm cases. Additionally providing low false alarm rates is as crucial as not missing real alarms since high false alarm rates may show the surveillance system unreliable which may cause operators to ignore real alarms.

A method using multi-camera system to detect unattended luggage in public areas when multiple objects exist in the FOV is proposed in [38]. In this study, objects are segmented using background modeling and objects' locations are extracted in order to handle the occlusions. Using heuristics based on the distance between the luggage and the owner of the luggage, the alarm is generated. Another approach which contains static objects detection, dynamic object detection, homographic transformation, height estimator and is capable of multiple objects tracking while using multiple cameras is presented in [39]. An alarm is given when one of the objects is stationary and the other one exit from the predefined position. A similar method analyzing object trajectories and presenting a real-time abandoned object detection algorithm is described in [40]. While detecting objects, a geometry based approach is utilized. A study that aims to detect packages that are left unattended in crowded environments is proposed in [41]. In this paper, firstly objects are tracked with a blob tracking approach driven on a trans-dimensional Markov Chain Monte

Carlo tracking. This method does not classify objects as people or luggage, only the result of tracking system is used to define abandoned object. In study [42], a left object detector which focuses on the merging and splitting of objects that is defined as a potential case to left the luggage unattended is proposed. By using background modeling, moving objects are tracked and at the same time a stationary object detector is executed. The stationary objects are associated with drop-off events and according to the distance between the owner of the object and the object left unattended, an alarm is given. Another abandoned object detection method which aims supporting an operator in guarding indoor environments, capable of classifying unattended objects as stolen or abandoned and finding the owner or the thief of the object is presented in [37]. In a different technique, objects are recognized based on their gradient histograms and trained using SVM to find the abandoned objects [43]. A study which copes with multiple occlusions, using multi-layer detection system for abandoned object detection is proposed in [44] and an unattended object detection algorithm which includes fusion of color and shape information of static foreground objects is presented in [45]. In [46], a study that uses long-term and short-term background modeling to detect abandoned object is described. In this method, connected components are classified as moving object, abandoned object, uncovered background or scene background according to pixel value changes in the short-term and long-term foreground images and an evidence image which only consist of the abandoned items is constructed. Although this method is very easy to implement and provide a new point of view to abandoned object detection task, since it does not discriminate or classify the objects as living or nonliving, it generates false alarms if a living object is stationary (For example people standing or sitting on a bench are detected as abandoned object). Additionally, this method detects nonliving objects as abandoned, even when the owner of nonliving objects stays next to the objects (which may cause false alarms).

CHAPTER 3

ADAPTIVE MEAN-SHIFT FOR AUTOMATED MULTI OBJECT TRACKING

One of the most popular tracking techniques is the mean shift algorithm due to its robustness, ease of implementation and computational efficiency. However, the standard mean shift has some drawbacks which affect tracking performance and cause inaccurate and/or false positive tracking. First of all, standard mean shift is not adaptable to changes in object size or shape. It requires correct kernel bandwidth selection to perform accurate tracking which is generally not applicable in the automated systems. Additionally, when the kernel shape does not fit the object, background information is included into the object model. Finally, the tracking might shift and even fail when object is occluded or background colors are similar to the foreground objects' colors.

A number of studies have attempted to solve these problems but they generally focus on a single shortcoming of mean-shift tracking. However, to achieve a robust tracking, all the drawbacks need to be handled. Besides, methods in the literature mostly require manual object initialization and need human interaction to define the objects that will be tracked. Furthermore, overwhelming majority of studies aim to track only one object at a time and does not provide a solution for tracking of multiple objects.

In this chapter, an easy to implement and fully automatic multiple object tracking algorithm based on standard mean shift method and called adaptive mean-shift for automated multi object tracking is proposed. This method is applied to visible band

images and thermal band images individually. An update mechanism is used to improve the tracking performance. Additionally, foreground object obtained from result of background subtraction is used to initialize and refresh trackers. For visible band, a shadow removal method is additionally used and false positives are decreased.

3.1. Method for Visible Band Image

In this section, the algorithm of adaptive mean-shift for automated multi object tracking and its application in visible band are given.

3.1.1. Overview

The block diagram of the proposed adaptive mean shift for automated multi object tracking for visible band is given in Figure 3.1. As shown in this figure, firstly background subtraction is applied to the visible image. Then, shadows are eliminated. Connected component analysis is used to classify objects as the ones which will be tracked or as noise which will be ignored. Finally, improved mean shift tracking which contains update condition, re-initialization of trackers, correspondence based object matching and standard mean shift tracking with masking the search area is applied and all the objects in the video sequence are tracked. These steps are described in more detail below.

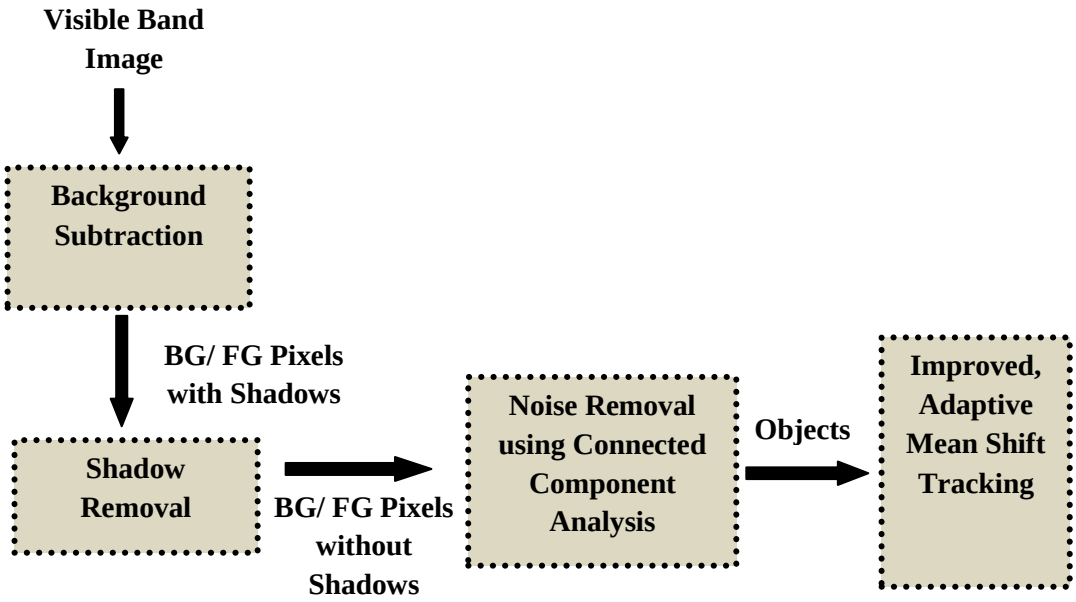


Figure 3.1: Block diagram of the adaptive mean-shift for automated multi object tracking for visible band

3.1.2. Background Subtraction

The object segmentation step is the initial step for object tracking and it directly affects the robustness and correctness of following steps. In order to correctly segment multiple objects in the scenario and decrease the number of false positives caused by object segmentation, a powerful background subtraction algorithm is needed. Therefore, in this thesis, improved, adaptive Gaussian mixture model reported in [6] as a reliable background modeling while being computationally not very complex has been decided to used.

This method is a pixel-based background subtraction method where a pixel belongs to the background or the foreground according to the following Bayesian rule R:

$$R = \frac{p(BG \mid \vec{x}(t))}{p(FG \mid \vec{x}(t))} = \frac{p(\vec{x}(t) \mid BG)\, p(BG)}{p(\vec{x}(t) \mid FG) p(FG)} \qquad (3.1)$$

where BG represents background while FG represents foreground object and $\vec{x}(t)$ is the value of pixel at time t. Probability of being foreground or background is equal

for a pixel, it is set as $p(BG) = p(FG)$, foreground object has a uniform distribution as $p(\vec{x}(t)|\, FG) = c_{FG}$ and pixel belongs to background when it is probability is greater than a threshold that discriminates background from foreground as it is shown in Eq. 3.2.

$$p(\vec{x}(t)|\, BG) > c_t = R_{c_{FG}} \tag{3.2}$$

when c_t is a threashold and $p(\vec{x}(t)|\, BG)$ is a background model.

In this background subtraction method, background modeling finds background pixels over a training set *X* and this model is updated with new samples in a predefined time to delete the old samples and adapt the algorithm to changes in background. For instance, if training set at time t is $X_t = \{x(t), \dots, x(t-T)\}$ when *T* represents time period X_t is updated with new samples at time *T* and the background model is refreshed according to new training data. The most important thing that should considered while refreshing background model is the possibility of a foreground object in previous background. Therefore, while estimating the new background model, the probability of a pixel is calculated with respect to background and the foreground object coming from previous X_t at time *T* as shown in Eq. 3.3.

$$New\ Background\ modeling = p(x(t)|\, X_t, BG + FG) \tag{3.3}$$

Each pixel is defined as a mixture of Gaussians with M components as follows:

$$p\left(x(t) \mid X_T, BG + FG\right) = \sum_{m=1}^{M} \hat{\pi}_m N\left(x; \mu_m, \hat{\sigma}_m^2 I\right) \tag{3.4}$$

where $\mu_1, \mu_2, \dots, \mu_M$ are mean values, $\sigma_1, \sigma_2, \dots, \sigma_M$ are variance values, $\pi_1, \pi_2, \dots, \pi_M$ are the weight values that are nonnegative and summation is one.

When new data sample *x(t)* at time t is given the Gaussian model parameters are updated using the Equations (3.5), (3.6), (3.7) [6].

$$\hat{\pi}_m \leftarrow \hat{\pi}_m + \alpha\left(o_m^{(t)} - \hat{\pi}_m\right) \tag{3.5}$$

$$\hat{\vec{\mu}}_m \leftarrow \hat{\vec{\mu}}_m + o_m^{(t)}(\alpha/\hat{\pi}_m)\vec{\delta}_m \tag{3.6}$$

$$\hat{\sigma}_m^2 \leftarrow \hat{\sigma}_m^2 + o_m^{(t)}(\alpha/\hat{\pi}_m)\left(\vec{\delta}_m^T\vec{\delta}_m - \hat{\sigma}_m^2\right) \tag{3.7}$$

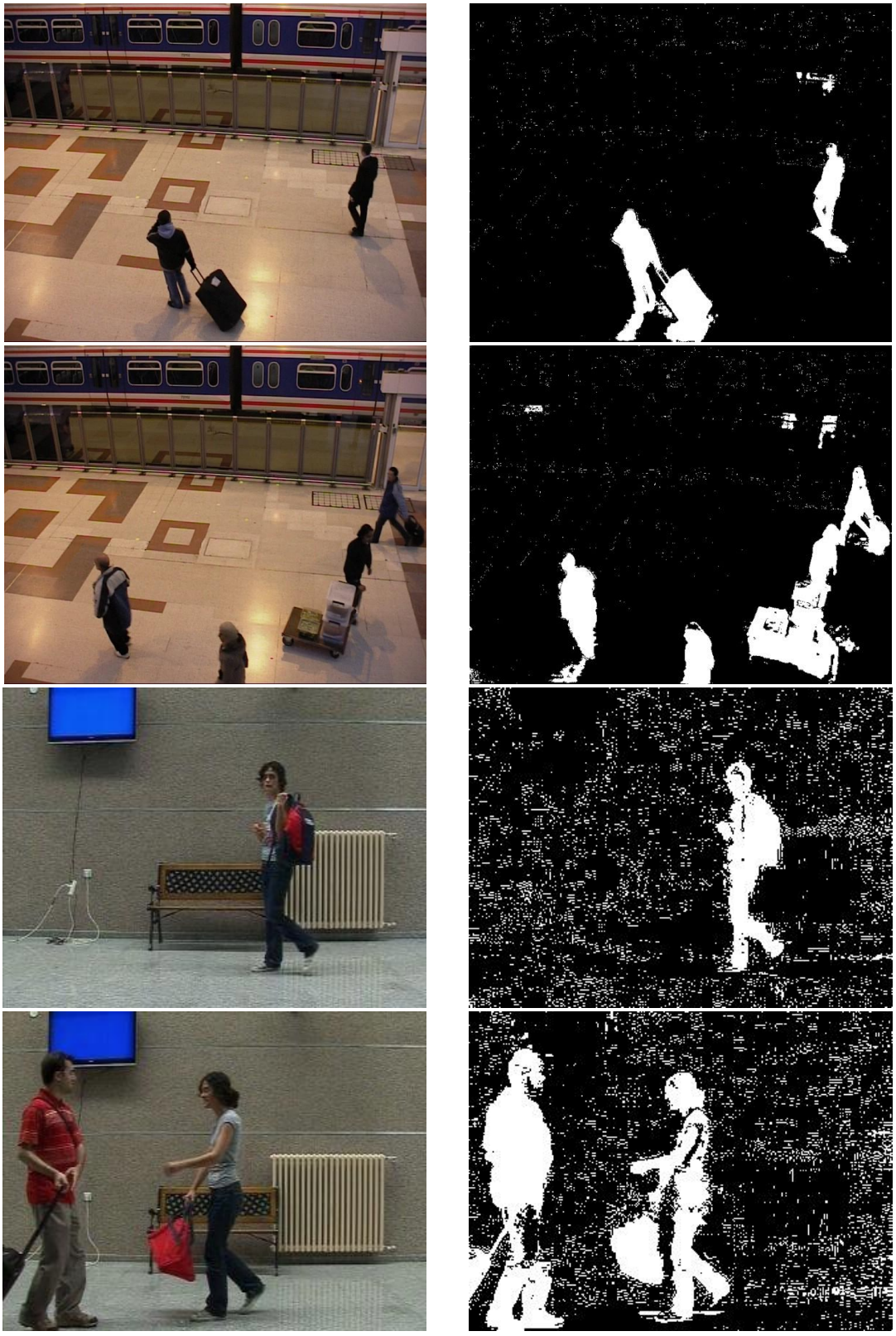
where $\delta_m = x(t) - \mu_m$, $o^{(t)}$ is ownership, and α is learning parameter, approximately $\alpha = 1/T$, T is a time period. For each component, m is set to one if it is close component to largest π_m and the others are set to 0. New sample is "close" to component if the Mahalanobis distance between them is less than four standard deviations. Square distance from mth component can be calculated by using Eq. (3.8).

$$D_m^2\left(\vec{x}^{(t)}\right) = \vec{\delta}_m^T\,\vec{\delta}_m / \hat{\sigma}_m^2 \tag{3.8}$$

where $D_m(x^{(t)})$ is the Mahalanobis distance from m_{th} component. If there is no close component, the parameters are reinitialized with $\pi_{M+1} = \alpha$, $\mu_{M+1} = x(t)$, $\sigma_{M+1} = \sigma_0$ where σ_0 is initial variance. Foreground objects are detected using the small weights π_M and B largest distributions with high π_M values are used to determine background model as follows:

$$B = \arg\min_b\left(\sum_{m=1}^{b}\hat{\pi}_m > (1 - c_f)\right) \tag{3.9}$$

where c_f is a measure of maximum portion of foreground objects' data. For instance, if the object becomes stationary for a while, since its π_M value becomes larger than c_f it is highly possible that this object will be detected as background in the next frames. The example results of this background subtraction method are given in Figure 3.2.



(a) (b)

Figure 3.2: (a) Visible band image, (b) Result of background subtraction

3.1.3. Shadow and Noise Removal

In tracking systems, shadows might cause problems which make the object tracking to perform inaccurately. For instance, they give rise to merging of objects, distortions of color histogram of objects and inclusion of background information due to the bounding box of the object become larger. An example result showing the tracking results while the shadow removal is not applied is given in Figure 3.3 and the need of shadow removal step is demonstrated.

Figure 3.3: An example tracking result when shadow removal is not applied

As seen Figure 3.3, bounding box of the object also includes the shadow and hence the tracker kernels are initialized with this larger bounding box. This causes background information to be included in the kernel and results in reduced tracking accuracy.

In this thesis, we use a shadow detection scheme which is based on the HSV color space and presented in [12]. A shadow detection method which is integrated into the Gaussian mixture model background subtraction could also be used. However, this technique [12] is chosen since shadow removal is contrived as an independent step which could be used with a different background subtraction algorithm such as minimum, maximum values or a median operator when proposed method is adapted

to another study. Additionally, it has been shown to be successful regardless of the applied background subtraction method [12].

We make use of the fact that shadow cast on the background does not significantly change its hue and saturation information while it decreases the saturation of the pixels. For each pixel which is detected as foreground, saturation, hue and value components are checked according to the Eq. (3.10) and the pixel is classified as a foreground pixel or shadow.

$$P(x,y) = \begin{cases} foreground\ if\ \ \propto \leq \frac{I^t(x,y).V}{B^t(px,y).V} \leq \beta\ \wedge\ |I^t(x,y).S - B^t(x,y).S| \leq\ T_s\ \wedge D_H \leq T_H\ ;\ \propto, \beta\ \in [0,1] \\ shadow\ otherwise \end{cases} \tag{3.10}$$

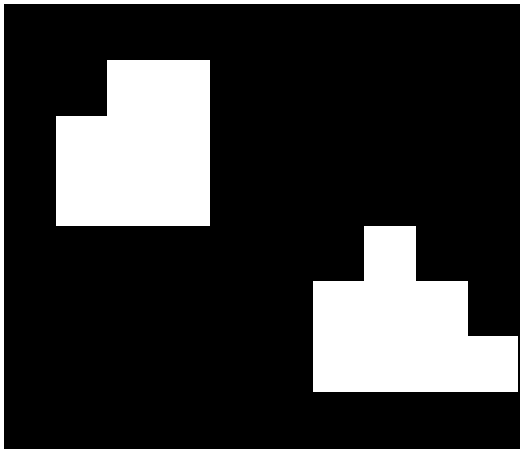
Where $I^t(x,y)$ is the pixel which is classified as foreground from the background subtraction step at time t and $B^t(x,y)$ is the background model pixel at time t. H denotes the hue component, S denotes the saturation component and V denotes the value component of a vector in the HSV space. α is a maximum value for the darkening effect of shadows on the background and β is the upper bound to handle the pixels that the background was darkened too little when compared to the effect of shadows. T_s is the threshold value that defines the upper bound of absolute difference between saturation of pixel and background model. T_H is defined as the upper bound of hue value [12].

After removing shadows, connected component analysis with 8-connectivity is applied. By using connected component analysis in addition to detecting and removing noise, the foreground objects which are aimed to be tracked are found.

In order to apply connected component analysis to a binary image, several algorithms exist in the literature [47, 48 and 49]. In this thesis, we applied a basic algorithm with 8-connected neighborhood which is proposed in [48]. According to this algorithm, the image is scanned until a foreground pixel is found. When a foreground pixel which was not labeled before is found, its coordinates are stored to check whether its

8 neighbors belong to a foreground object or not. If any neighbor is a part of a foreground object, then the same label is given to that pixel. This procedure continues until there is no pixel which is not labeled.

In Figure 3.4, an example image which a connected component analysis can be applied is shown. In this figure, white pixels represent the foreground pixels while black pixels represent background pixels.



0	0	0	0	0	0	0	0	0	0
0	0	1	1	0	0	0	0	0	0
0	1	1	1	0	0	0	0	0	0
0	1	1	1	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	1	1	1	0
0	0	0	0	0	0	1	1	1	1
0	0	0	0	0	0	0	0	0	0

(a) (b)

Figure 3.4: Connected component analysis in binary image (a) Foreground image, foreground pixels are colored white, background pixels are colored black, (b) Connected component labeling matrix

After each pixel is labeled, to eliminate the noise, the bounding box of the object (which is calculated by searching each pixel to find the minimum, maximum x coordinate and the minimum, maximum y coordinate that the object has), the number of pixels that each connected component has and the area of the object's bounding box are found. Then, density of each object is calculated by using Eq. 3.11.

$$D = N / A_{rect} \tag{3.11}$$

Where D is the density of object, N is the number of pixels that object has, A_{rect} is the area of the bounding rectangle.

After finding the density of the object, each connected component is classified as object that will be tracked if its density is greater than the density threshold and number of pixels that belongs to this object is greater than the maximum number of pixel threshold, otherwise the connected component is classified as noise and it is ignored.

Figure 3.5 shows the example shadow and noise removal results. In these images shadows are shown in green and foreground pixels are shown in red.

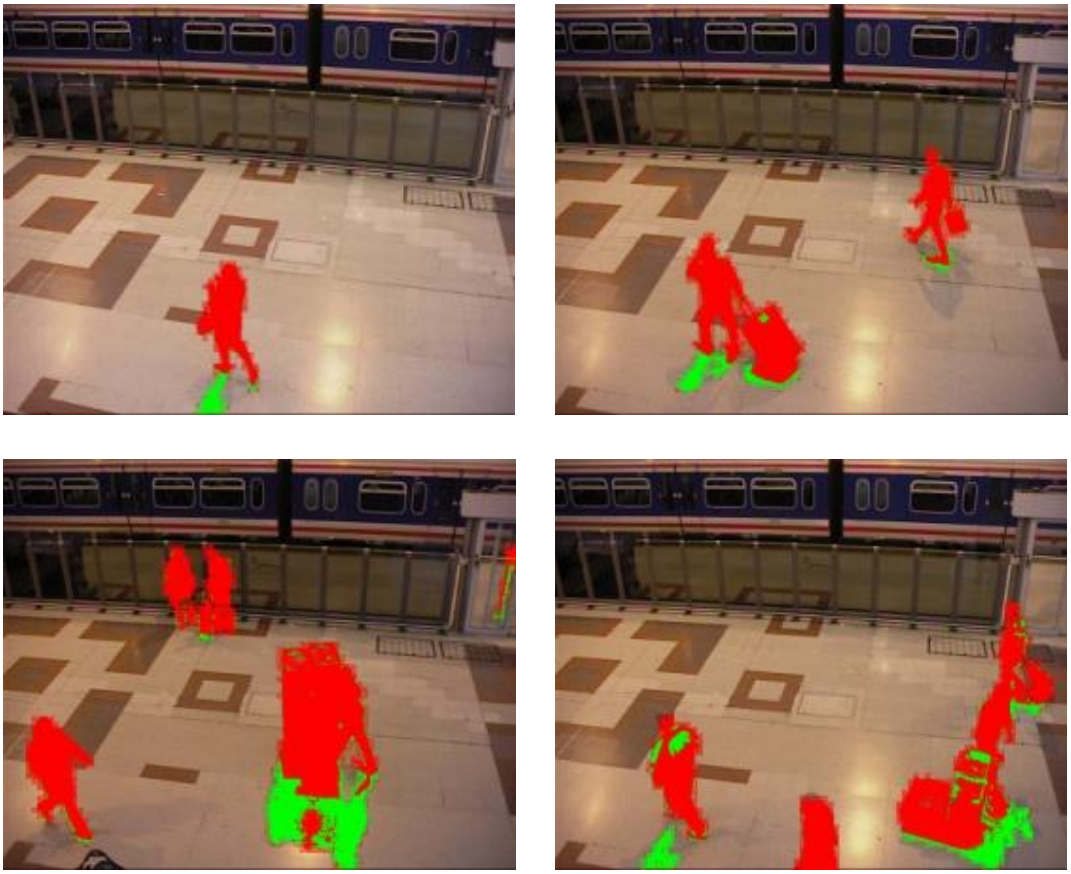


Figure 3.5: Result of shadow and noise removal.

3.1.4. Standard Mean-Shift Tracking Algorithm

The mean shift tracking method is an iterative method based on object representation. Additionally, it is an optimization problem and uses a nonparametric kernel. It basically tries to find an object in the next image frame which is most similar to the initialized object (object model) in the current frame. Similarity is found by comparing the histogram of the object model and histogram of the candidate object in the next frame.

At the initialization step, an object model which is aimed to be tracked is selected. Bin size, kernel function, size of the kernel and maximum iteration number are determined. Color histogram of the object model is found and the probability density function (pdf) of the object model is calculated as follows:

$$q_u = C \sum_{i=1}^{n} k\,(||x_i||^2)\delta[b(x_i) - u] \tag{3.12}$$

In this equation, *k* is the kernel function which gives more weight to the pixels at the center of the model, *C* is a normalizing constant which provides that sum of the histogram elements is 1, *u* represents histogram bin and *n* is the number of pixel in the object model. δ is the Kronecker delta function and *b* represents histogram binning function for pixels at location x_i [28].

After defining the target model in initialization step, candidate model is constructed. Similar to the target model's pdf, candidate model's pdf at location *y* is given by

$$p_u(y) = C_h \sum_{i=1}^{n_h} k\left(\left|\left|\frac{y - x_i}{h}\right|\right|^2 \delta[b(x_i) - u]\right) \tag{3.13}$$

Where *h* is the kernel size which provides the size of the candidate objects [28].

After defining the pdf of candidate model, it is compared with target model's pdf. To compare color based pdfs, generally the metric derived from Bhattacharyya coefficients is used. In Eq. 3.14 $p(u)$ and *q(u)* represent the Bhattacharyya coefficients.

$$\rho[q_u, p_u(y)] = \sum_{n=1}^{m} \sqrt{p_u(y) q_u} \tag{3.14}$$

The larger ρ means the more similar the pdfs are. If the candidate model is not similar to the target model then the current search area is shifted. This iteration continues until the result of similarity is less than a threshold or when the numbers of iterations reach a predefined number. By applying this method to each video frame, object model can be tracked over time.

3.1.5. Improved, Adaptive Mean Shift Tracking Algorithm

Object detection which could be either manual or automatic is the main step for object tracking. In the literature, many studies use manual object detection and initialization initiated by an operator. However, if manual initialization is used, since new objects could not be tracked when they enter the scene after the initialization frame, it is expected that all objects exist in the starting frame of the video sequences or the operator regularly detects all the new objects. On the other hand, when automatic initialization is applied, any new object entering the scene could be tracked without any input from human operator.

In this study, foreground detection which is obtained by background subtraction is used for automatic initialization. Firstly, improved adaptive Gaussian background subtraction (Section 3.1.2) is applied, then noise and shadow removal methods (Section 3.1.3) are executed, connected component analysis is used and objects that will be tracked are determined. In addition to the benefits of using the foreground detection for automatic initialization of objects, it is also used as a mask to decrease the search area of the mean shift tracker. This increased our system's tracking accuracy and performance since there is no need to search the entire frame. Additionally, the required number of iterations to find the new position of object model is decreased.

Even though tracking objects by only using the result of foreground detection seems possible, it is not a robust method. When multiple objects are required to be tracked in crowded places and in the presence of occlusions, matching of objects and finding

the correspondences become difficult. To solve this problem, in addition to information coming from foreground detection, correspondence based tracking similar to the one proposed in [22] can be used. However, applying this method in every frame is not efficient since it has high memory requirement to hold the objects and also objects' correspondence objects.

Although using foreground detection in the tracker initialization step has advantages, it is not sufficient to make the system fully automatic since it still does not detect the new objects entering the scene or the objects leaving the scene. To solve this and to have a system which is adaptable to changes in objects' size and shape and to handle inclusion of background information, we reinitialize the trackers by using foreground information at regular time intervals. This update mechanism which includes re-initialization of trackers, an update condition, standard mean shift tracking with masking the search area by object's boundary and correspondence based object matching is shown in Figure 3.6.

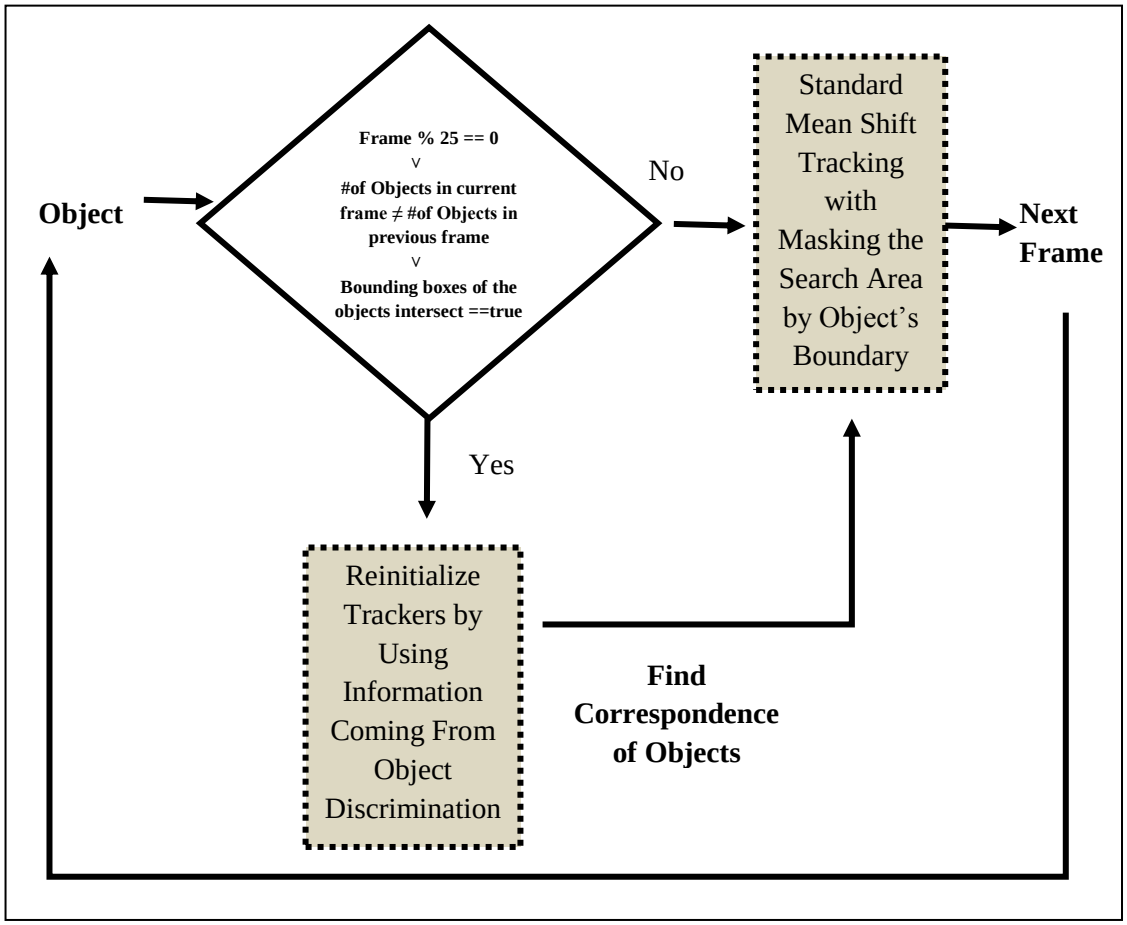


Figure 3.6: Improved, adaptive mean shift tracking

To handle the change in size or shape we update trackers every 25 frames. To detect new objects as well as objects that leave the scene, numbers of objects in subsequent frames are compared and if those numbers are not equal then trackers are updated. To handle occlusion and split and to detect newly emerging objects; the location of bounding box of all objects are compared. If an intersection exists then trackers are refreshed to handle inclusion of front objects color. However, as a result of re-initialization, trajectories of objects (object correspondence) are lost. To overcome this, we also find the correspondence of objects after each re-initialization step.

To establish the matching between objects and provide the correspondence in frames, we adapted correspondence based tracking method [22] to our method and used object's size, center of mass, bounding box and color histogram features. Firstly, we check whether an object O_i is close and similar to an object O_p in previous frame.

Closeness is defined as the distance between the center of mass of these two objects (O_i and O_p) and Euclidean formula (Eq. 3.15) is used to calculate this distance. Similarity is calculated by using size ratio of the objects (Eq. 3.16). If distance between object O_i and object O_p is smaller than a distance threshold and size ratio of objects O_i and O_p is smaller than a size threshold, we define object O_i as corresponding object for O_p. Using closeness was a successful criterion since the displacement of an object between adjacent frames should be small. However, it is not a sufficient criterion since objects that are close to each other in previous frame may interfere and the matching can be done wrongly. Using similarity criterion is also useful as objects do not scale too many between consecutive frames [22].

$$d(O_p, O_i) = \sqrt{\sum_{i=1}^{2} (x_{ip} - x_{io})^2} \quad) \leq t_{distance} \tag{3.15}$$

Where $d(O_p, O_i)$ is Euclidean distance, *x* represents x components of center of mass of objects O_i and O_p when *i* is equal to one and it represents y components of center of mass of objects when *i* is equal to two. $t_{distance}$ is distance threshold.

$$if\ S_p > S_i\ ,\ \frac{S_p}{S_i} \leq t_{size} \quad otherwise\ \frac{S_i}{S_p} \leq t_{size} \tag{3.16}$$

Where S_p is size of object O_p and S_i is size of O_i.

If object O_i is not a match to object O_p then there are two possibilities: O_i could be a new object or it could have been occluded by another object. To check whether an occlusion exists or not, we first compare the number of objects, if there is a decrease in number of objects then an occlusion is possible. If object O_i's bounding box is overlapping with bounding boxes of O_p and O_t then it is highly possible that O_p and O_t are occluded to generate object O_i (Figure 3.7). In such a case a tracker is created to follow O_i and colour histograms of O_p and O_t are stored in order to compare when a split occurs.

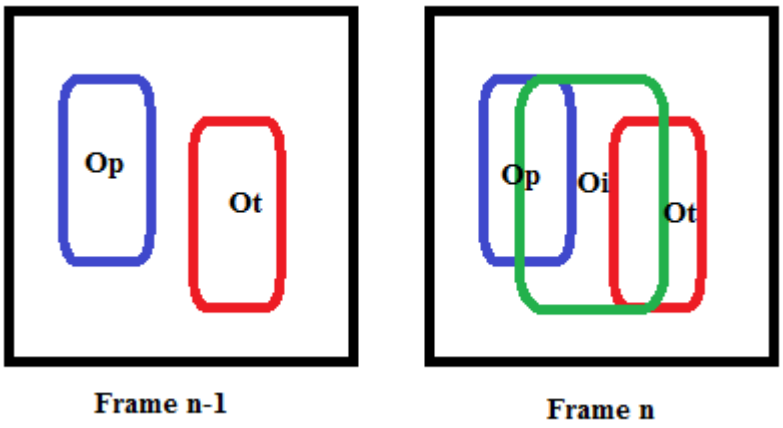


Figure 3.7: Occlusion detection [22]

Except these, for each object that enters the scene, we check whether its bounding box is overlapping with an occluded object or not. If its bounding box intersects with an occluded object then we compare its histogram pdf with occluded object's histogram pdf in order to handle a possible split. To compare histogram pdfs we used Bhattacharyya coefficients (Eq. 3.14), similar to the mean shift tracking. If there is a similarity, in other words if the distance between pdfs are smaller than a threshold, then object is matched with the occluded object and that occluded object's histogram is removed.

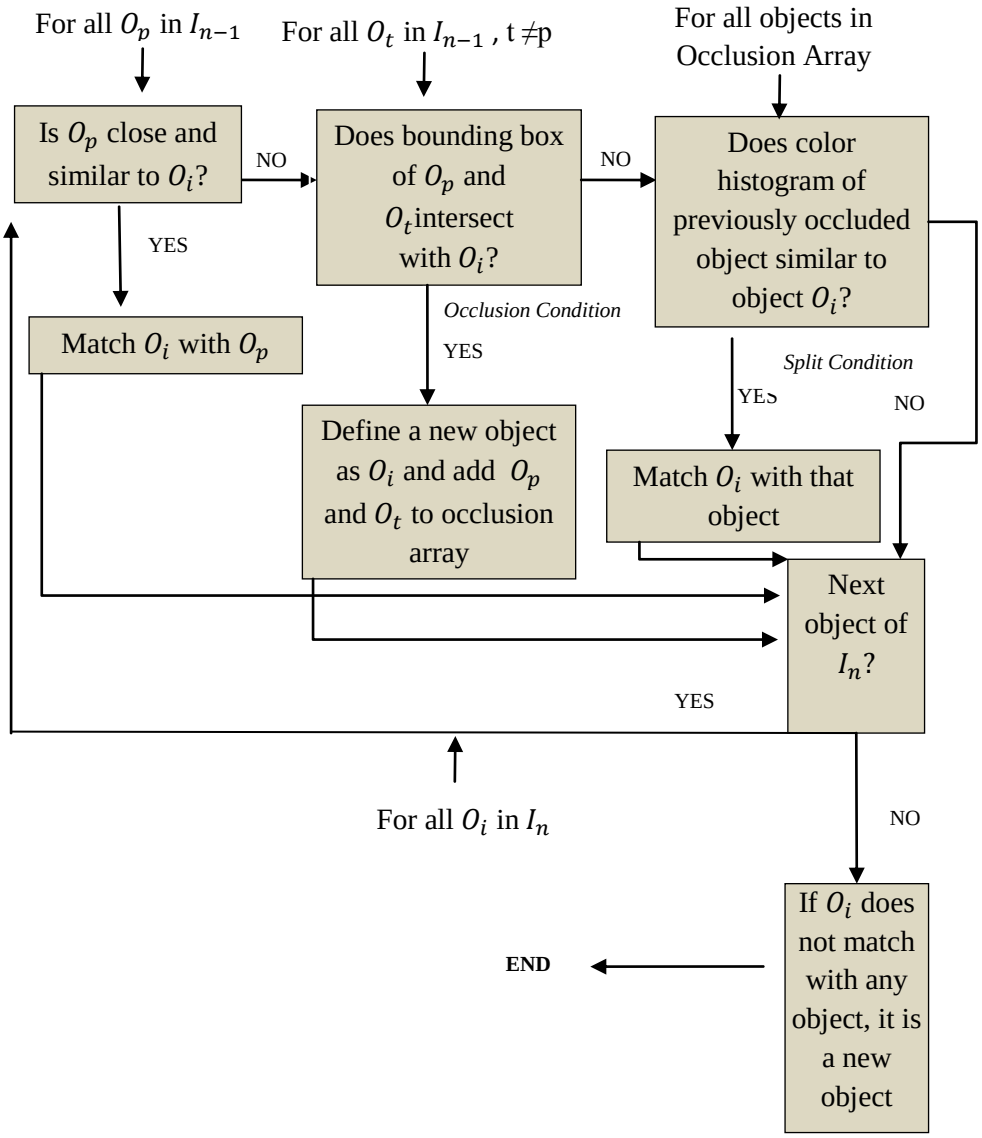


Figure 3.8: The correspondence based object matching after re-initialization of trackers

If an object does not match to an object O_p and occlusion, split do not exist, then this object is assumed to be a new object and a new tracker is defined to track it. The mechanism to establish the matching of objects when a re-initialization occurs is given in Figure 3.8.

Example result of adaptive mean shift for automated multi object tracking in visible domain is given in Figure 3.9.

Figure 3.9: An example result of adaptive mean shift for automated multi object tracking in visible domain

3.2. Method for Thermal Band Image

In this section, the algorithm of adaptive mean-shift for automated multi object tracking and its application in thermal domain is given.

3.2.1. Overview

The block diagram of the adaptive mean shift for automated multi object tracking in thermal domain is given in Figure 3.10. As it is done to in the visible band, firstly background subtraction is applied to the thermal image. Then connected component analysis is used to determine objects which will be tracked and camera noise is ignored similar to the implementation in visible domain. Finally, improved mean shift tracking is applied and all the objects in the video sequence are tracked. These steps are described in more detail below.

As it is seen in Figure 3.10, the main difference in application of adaptive mean shift for automated multi object tracking between visible and thermal domain is the shadow removal step which is applied in visible band while not applied in thermal image. However, similar to shadows causing problems in the visible band, thermal

reflections in thermal domain may cause inclusion of background information to the object model and shifts in object tracker.

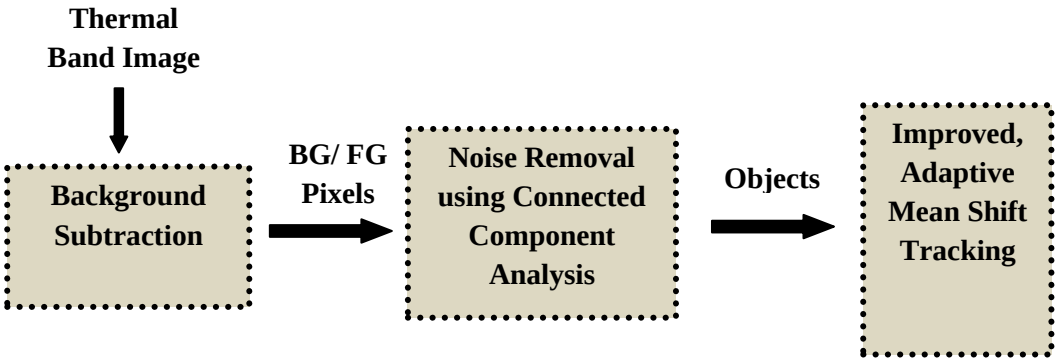


Figure 3.10: Block diagram of the adaptive mean-shift for automated multi object tracking for thermal band

3.2.2. Background Subtraction and Noise Removal

Background subtraction is applied as mentioned in Section 3.1.2 and foreground objects are segmented. Example results of this step are given in Figure 3.11. As it is seen, thermal reflection is also segmented from result of background subtraction in thermal image which may cause false positives and also inclusion of background information to the object model and which result in shift in object tracker. After background subtraction is applied, the camera noise is removed utilizing connected component analyses as explained in Section 3.1.3.

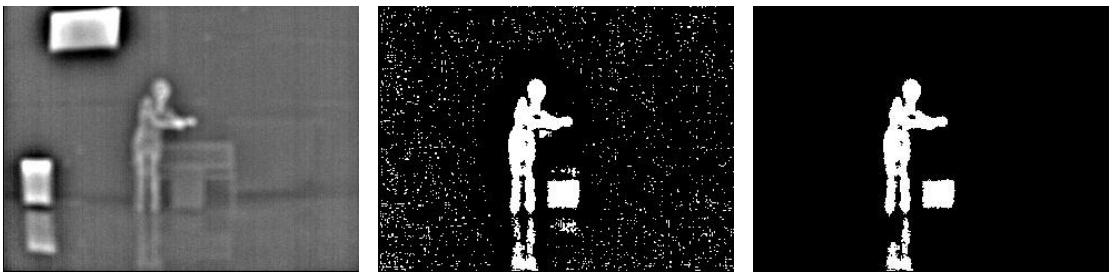


Figure 3.11 (continued)

(a) (b) (c)

Figure 3.11: (a) Thermal image, (b) Result of improved adaptive Gaussian mixture model for background subtraction in thermal band image (c) Result of noise removal

3.2.3. Improved, Adaptive Mean Shift Tracking Algorithm

Improved adaptive mean shift tracking is applied similarly as explained in 3.1.5 and multiple objects in thermal domain are tracked. Differently, in this step one dimensional (1D) histogram which contains infrared band (I) is constructed while YCrCb color space is used in visible domain. Additionally, for occlusion and split condition (Figure 3.8) instead of comparing the color histograms' of merged objects to matched the objects when a split occur, the sizes' (although it is not a robust feature as color histogram) of these objects are compared.

Example result of adaptive mean shift for automated multi object tracking in visible domain is given in Figure 3.12.

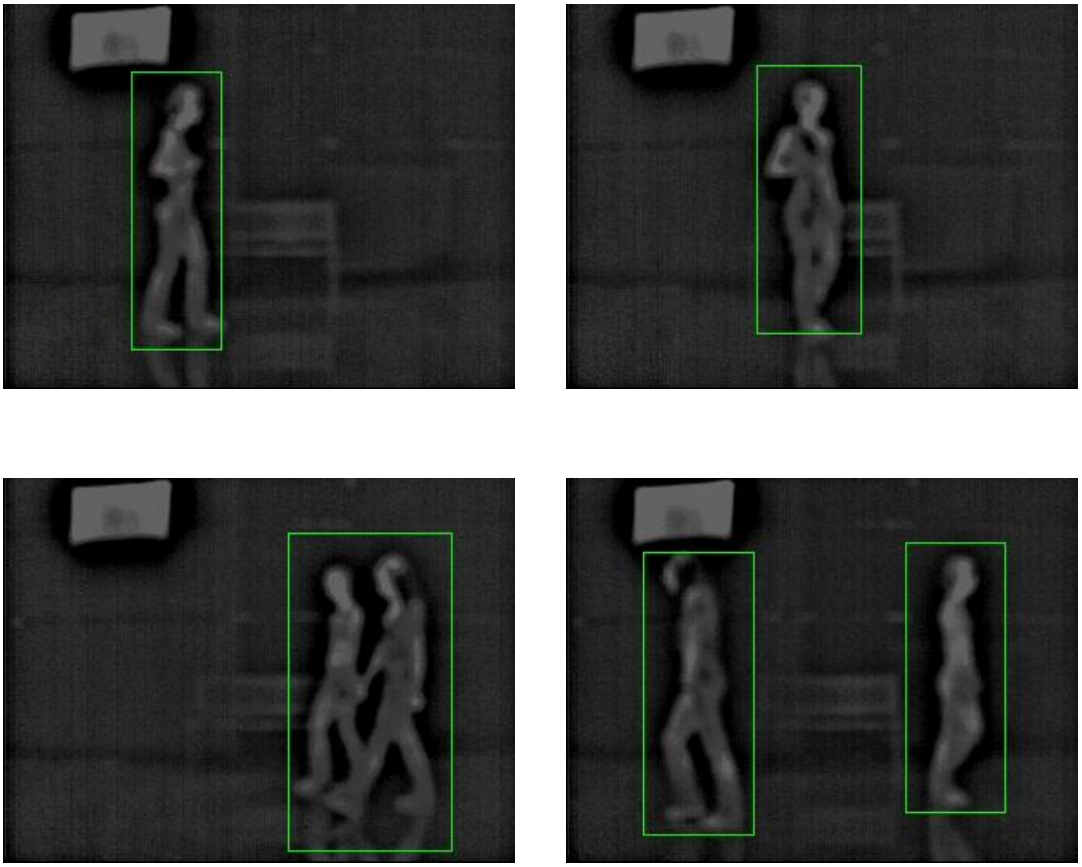


Figure 3.12: Example results of adaptive mean shift for automated multi object tracking in thermal domain

CHAPTER 4

A MULTIMODAL APPROACH FOR INDIVIDUAL TRACKING OF PEOPLE AND THEIR BELONGINGS

In this chapter, a fully automatic surveillance system for indoor environments which is capable of individual tracking of multiple objects (such as people and their belongings) and an abandoned object detection algorithm are proposed.

Multiple objects are tracked using the proposed improved adaptive mean shift tracking in Section 3.1.5. In addition to visible band, thermal images are also used and these two modalities are fused to track people and the objects they carry separately using their heat signatures which allow making higher level inferences and make timely detection of potential treats such as abandoning of an object by a person. By using the information coming from different modalities, trajectories of the living and nonliving objects are found, owner of the nonliving object is determined and abandoned objects are detected. Better tracking performance is also achieved compared to using single modality. Adaptive background modeling (Section 3.1.2) and local intensity operation (LIO) defined below are used in association with mean-shift tracking for fully automatic tracking. Trackers are refreshed as explained in Section 3.1.5 to resolve the possible problems which may occur due to the changes in object's size, shape and to handle occlusion, split and to detect newly emerging objects as well as objects that leave the scene.

4.1. Overview

Individual tracking of objects such as people and the luggage they carry is important for video surveillance applications as it would provide making timely detection of potential threats, such as abandoning of an object by a person. Additionally, by using the trajectories of these living and nonliving objects, interactions between them could be extracted. However, this is a challenging problem and in the literature, people and objects they carry are tracked as a single object.

The main steps of the proposed method are illustrated in Figure 4.1. As shown in this figure, firstly background subtraction is applied to the visible band image. Then, connected component analysis is utilized to remove the noise. LIO is applied to the thermal image and the result of this operation is post-processed to complete and close possible holes which might be formed after local intensity operation. Object discrimination step is the fusion step which uses both modalities. We apply fusion at this stage as in a recent study, it has been reported that fusion after object detection approach is the most successful scheme [14]. After this step, a rule based method and connected component analysis are used to extract objects and classify them as living or nonliving. Finally, each object (living and/or nonliving) are tracked using our improved, adaptive mean shift tracking algorithm (Section 3.1.5). While tracking objects, living and nonliving objects are also associated with each other and owner/carried object relation is set for tracked objects. Abandoning of an object is detected by using these relations and tracking the objects separately. These steps are described in more detail below.

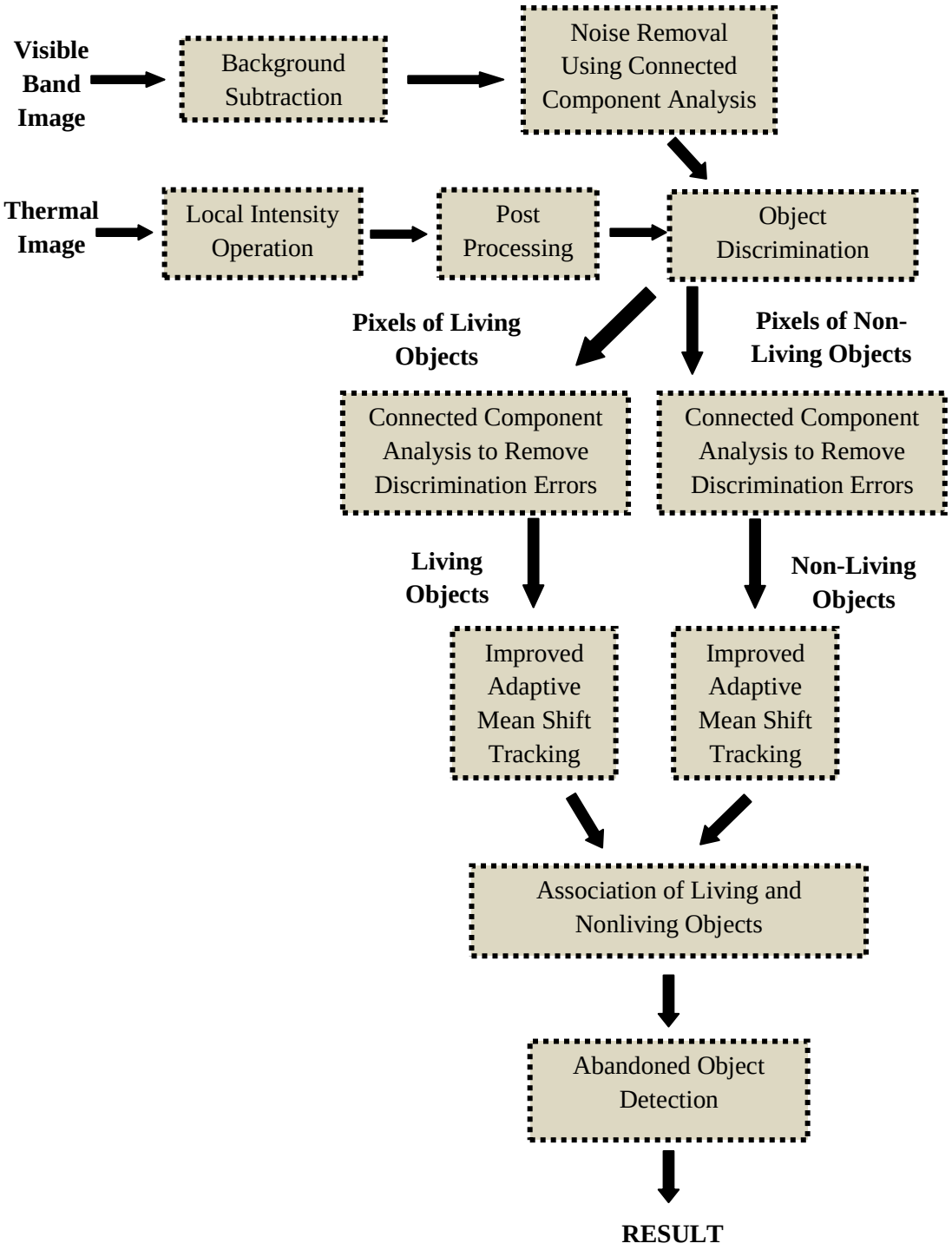


Figure 4.1: Block diagram of the individual tracking of people and their belongings and abandoned object detection

4.2. Background Subtraction and Noise Removal

Improved adaptive Gaussian mixture model background subtraction is applied to visible band images as mentioned in Section 3.1.2 and foreground objects are segmented. Then, the camera noise which cause false positives in subsequent steps is removed utilizing connected component analyses as explained in Section 3.1.3. Example result of this step is given in Figure 4.2.

(a) (b) (c)

Figure 4.2: (a) Visible band image (b) Result of background subtraction (c) Result of noise removal

4.3. Local Intensity Operation (LIO)

To discriminate people and their belongings, heat signature information is used. Since thermal images are constructed from emitted energy and living objects emit more energy compared to nonliving objects, pixels of living objects appear brighter than pixels of nonliving objects (in white-hot setting). The study in [50] uses LIO which makes brighter the bright pixels while makes darker the dark pixels in order to detect defects in thermal images. We also utilized this operation to segment pixels of living objects.

Before applying this method to thermal band images, if the contrasts of these images are low, a contrast stretching algorithm could be applied to segment living objects more accurately. As our test images have sufficient contrast, we have omitted this step.

According to this method, $I(x,y)$ is given as a pixel in thermal image written as z_0, and neighbors of it $I(x\text{-}1,y\text{-}1)$, $I(x\text{-}1,y)$, $I(x\text{-}1,y\text{+}1)$, $I(x,y\text{-}1)$, $I(x,y\text{+}1)$, $I(x\text{+}1,y\text{-}1)$, $I(x\text{+}1,y)$, $I(x\text{+}1,y\text{+}1)$ are written as z_1, z_2, z_3, z_4, z_5, z_6, z_7, z_8 respectively. Then, Z will be product of the intensity value of neighboring pixels:

$$Z = \prod_{k=0}^{8} z_k \tag{4.1}$$

An image is created according to Z for each pixel in thermal image by defining intensity brightness operation using Eq. (4.2) where $g(x,y)$ is the pixel value at location (x,y).

$$g(x,y) = Z \tag{4.2}$$

After that, these image's pixels are normalized to gray-scale range by diving the pixels to the maximum pixel value. Then Mean Absolute Thresholding (MAT) which is a kind of segmentation method and calculates as Eq. 4.3 is applied to get better results.

$$T = round\left(\frac{I_{max} - I_{min}}{2}\right) \tag{4.3}$$

Where T is the threshold, I_{max} is the maximum intensity, I_{min} is the minimum intensity.

As an alternative to multiplication of the intensity value of neighboring pixels (Eq. 4.1), addition of the intensity value of neighboring pixels (Eq 4.4) could be applied.

$$Z = \sum_{k=0}^{8} z_k \tag{4.4}$$

Examples of algorithms' results are shown in Figure 4.3. As it is seen from the results, the segmentation result of LIO operation when Eq. 4.1 is applied segments the living objects more successfully compared to Eq.4.4 while producing more noise. However, this noise could be eliminated by the following post processing operation. As a result, we have decided to use multiplication in LIO.

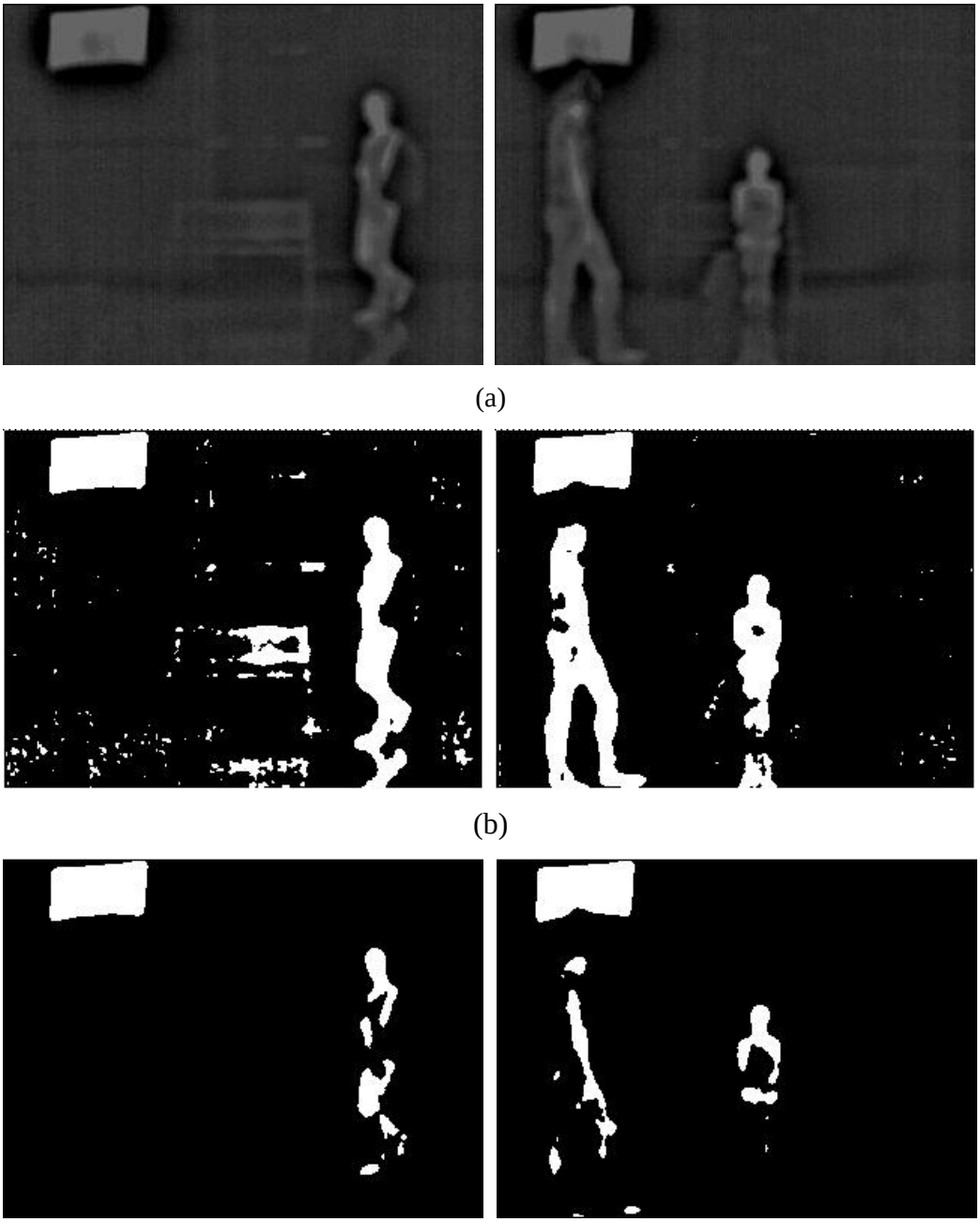


Figure 4.3: (a) Two unprocessed thermal images, (b) Segmentation results for these images using LIO with multiplication, (c) Segmentation results for these images using LIO with addition

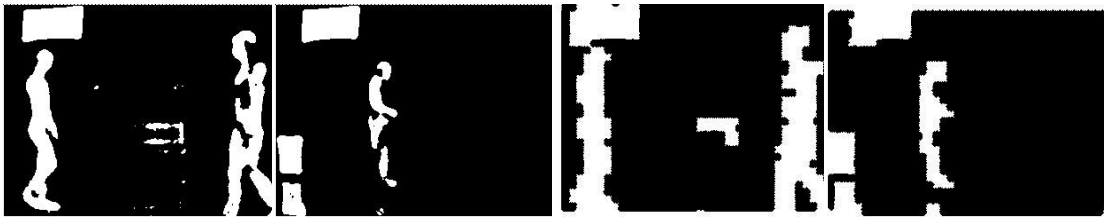
It has to be noted that, besides the living objects, hot objects such as heating systems, television screens (which have been used in airports to track the departures and/or arrivals) or any nonliving objects radiators (as seen in Figure 4.3) which are

hotter than the environment are also captured brighter than the other objects and segmented as a result of this process. However, as will be explained in later sections, since such objects belong to the background in visible image, our method does not discriminate these objects as people and, false alarms due to stationary hot objects are prevented.

Additionally, in order to overcome possible problems which may occur because of the changes of pictures in television screen, a mask which excludes the area of television screen in the image could be used.

4.4. Post Processing

The algorithm explained in Section 4.3 may not find the object precisely and some gaps may be observed on the object's body due to the clothing as seen in Figure 4.4. These problems are rectified with post-processing. To make these objects single piece, it is needed to complete and close these holes in binary images by using some morphological operations. First, objects in binary images (result of LIO, Figure 4.3b) are completed by hole-filling using connected component analyses explained in Section 3.1.3. Then, these binary objects are closed to eliminate small holes. Example results of post-processing step are shown in Figure 4.4.



(a) (b)

Figure 4.4: (a) Segmentation results for images using LIO with multiplication, (b) Post-processed images

4.5. Object Discrimination

Object discrimination step is the fusion step where both thermal and visible band images are used. It is the main step for individual tracking of objects such as people and their belongings.

In this step, the objects coming from result of background subtraction and noise removal (a binary image) in visible data (Section 4.2) and the objects coming from result of LIO and post processing (a binary image) in thermal data (Section 4.3 and 4.4) are utilized.

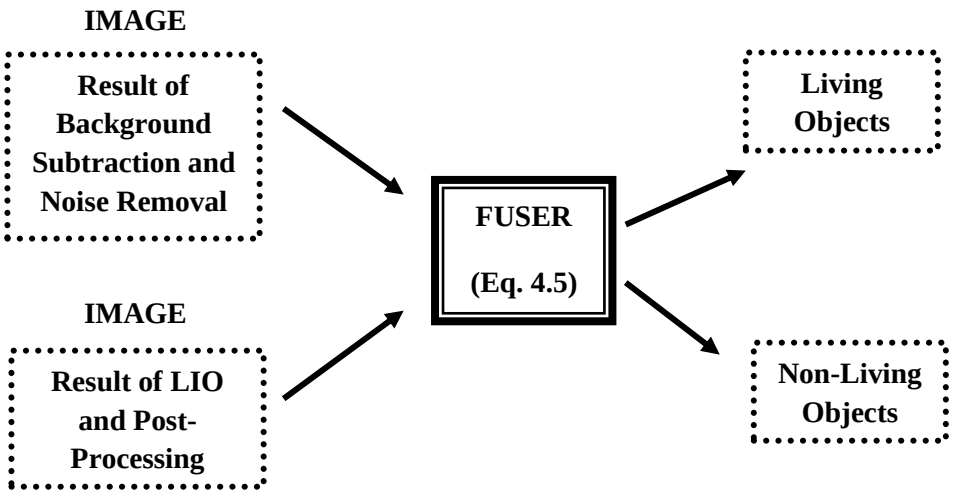
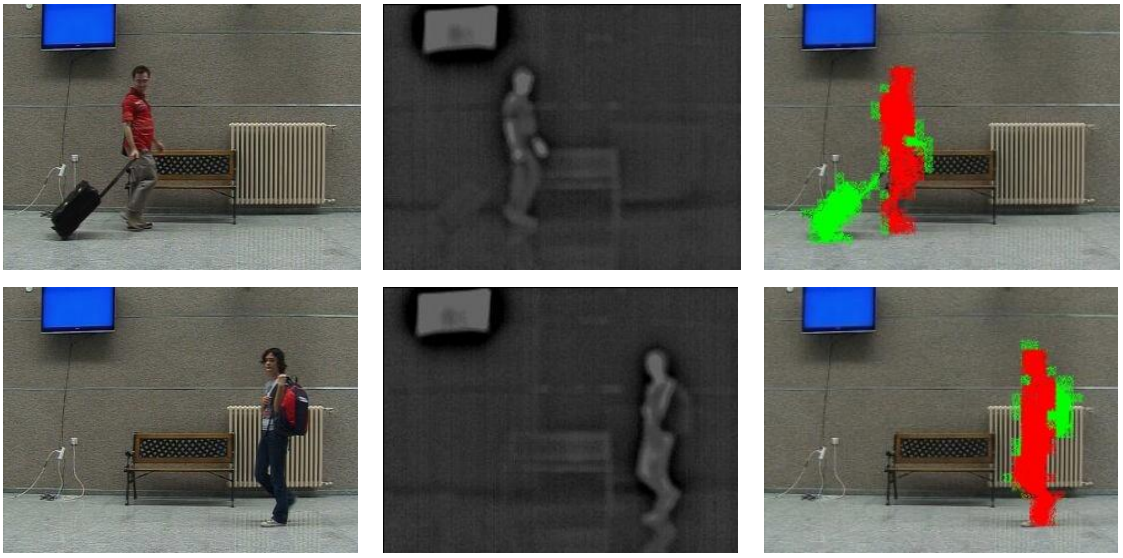


Figure 4.5: Object discrimination

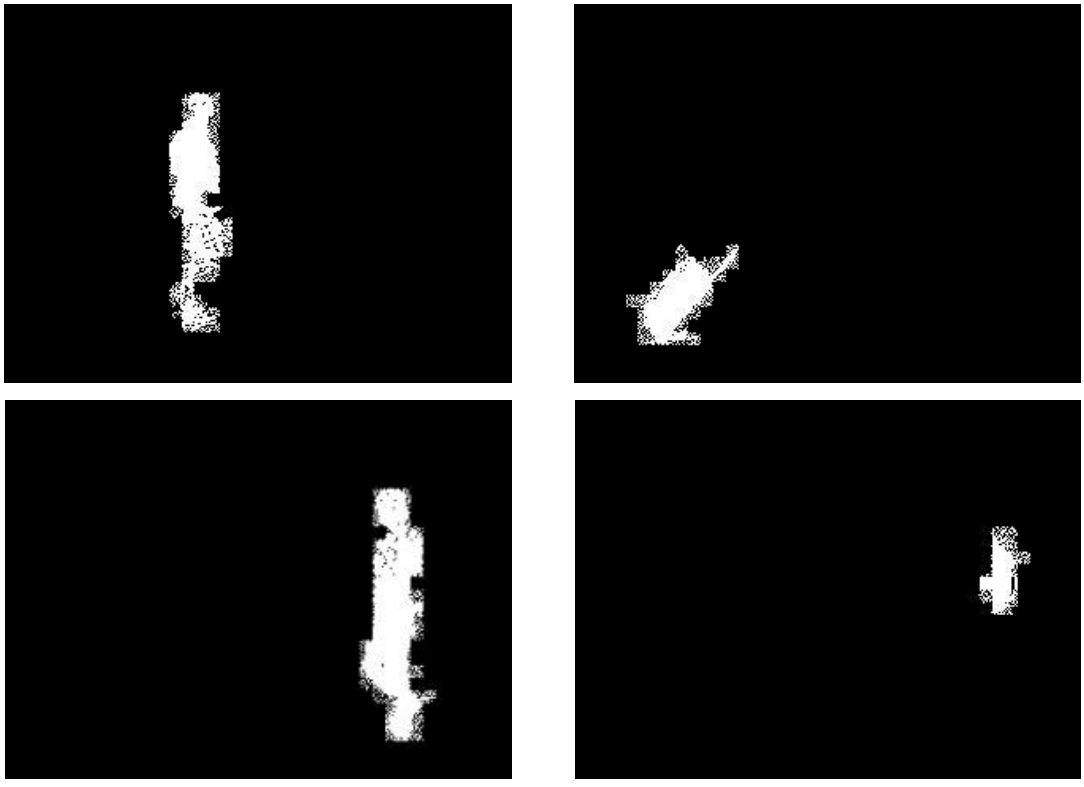
Using the rule given in Eq. 4.5 and connected component analysis, objects are extracted and classified as living (people) or nonliving (belongings) as shown in Figure 4.5.

$$F(x,y) = \begin{cases} living\ object\ (people), & R_V(x,y) \neq 0 \wedge R_T(x,y) \neq 0 \\ nonliving\ object\ (belonging), & R_V(x,y) \neq 0 \wedge R_T(x,y) = 0 \end{cases} \quad (4.5)$$

where $F(x, y)$ is the fusion result, $R_V(x, y)$ is the pixel value after background subtraction and noise removal in visible data and $R_T(x, y)$ is the pixel value after local intensity operation and post processing in thermal domain. By using this rule, thermal reflection and hot objects such as heating systems and radiators are not classified as living objects and possible false alarms are prevented.



(a) (b) (c)



(d) (e)

Figure 4.6: Results for two set of images (a) Visible band images, (b) Corresponding thermal images, (c) Object discrimination results with error, (d) Segmented living object, (e) Segmented non-living object.

After this operation some discrimination errors may occur especially around the living object due to inaccuracies in registration of thermal and visible band images. To handle these errors, same method which is presented in Section 3.1.3 to remove noise is applied and errors are eliminated. Example results of this step are shown in Figure 4.6. The nonliving object is shown in green and the living object is shown in red.

4.6. Improved, Adaptive Mean Shift Tracking Algorithm

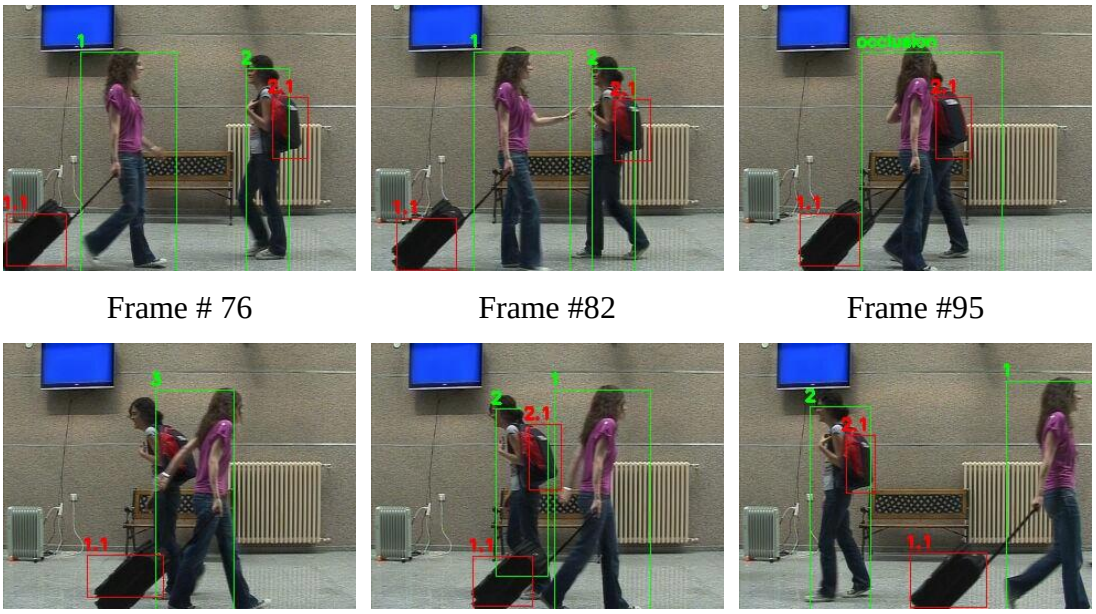
After objects are discriminated as living or nonliving, in order to track these objects, the algorithm explained in Section 3.1.5 is used. However, instead of using foreground detection, since we have two modalities, the result of object discrimination step is used.

By using the information coming from object discrimination step, the system becomes fully automatic and bounding box information is used as a mask to decrease the search area of the living and nonliving objects' mean shift tracker. By this way, the tracking accuracy and the required number of iterations to find the new location of living and nonliving objects are decreased. New objects (living or nonliving) entering to the FOV and objects that are leaving the scene could all be detected immediately. By regularly updating the trackers, mean shift becomes more adaptable to changes in object size and shape. Occlusion, split and merging scenarios which are not handled by standard mean shift are also handled.

4.7. Association of Living and Nonliving Objects

While tracking objects, ownership information is extracted and living (people) and nonliving objects (belongings) are associated with each other. To find the ownership, the closeness criterion is used. To calculate the distance between each nonliving object and living object, Euclidean distance is utilized (Eq. 3.15) and the nonliving object is assigned to the nearest living object. While determining the ownership, it is assumed that object is not handover to another person. Therefore, once nonliving

object is appointed to a living object, it is no longer appointed to another living object. In other words, if a nonliving object is exchanged, the index of nonliving object (which is given by associating the nonliving object with a living object) remains as it is initially assigned. On the other hand, if a nonliving object is associated with a living object while that living object is occluded by another living object since a wrong association is strongly possible, it is assigned to the living object after split occurs. If these objects form a new object by merging, then in the next update nonliving object is associated with the merged object. Example results of this association step are given in Figure 4.7.



Frame # 76 Frame #82 Frame #95

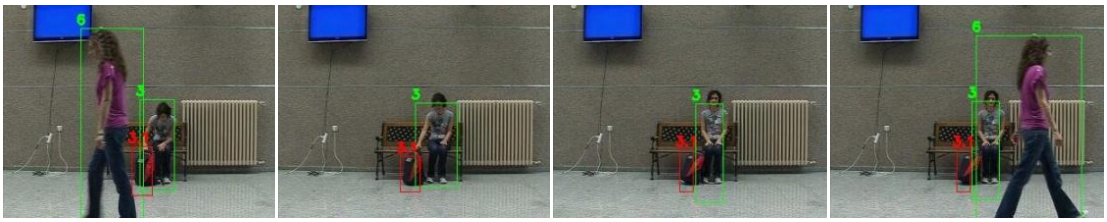
Frame #106 Frame #109 Frame #125

Figure 4.7: Association of objects with their owners

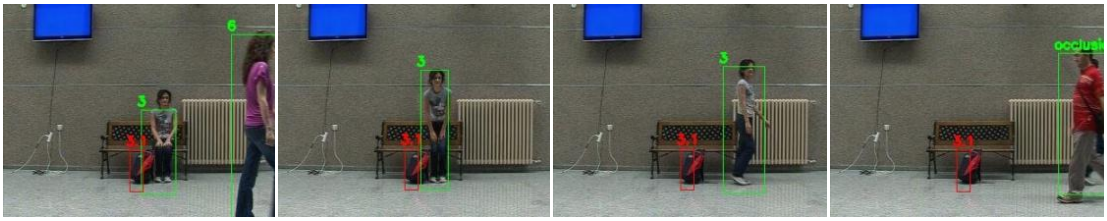
As it is seen in Figure 4.7, bounding boxes of nonliving objects are shown in red and bounding boxes of living objects are shown in green. To denote the association of an object with a person, nonliving object is indexed with a notation (Owner Index.Object Index For The Owner). The number before dot shows the nonliving object's owner's index and the number after the dot shows the index of nonliving object. For example in Figure 4.7, 1.1 is the object belonging to person 1 and 2.1 is the object belonging to person 2.

4.8. Abandoned Object Detection

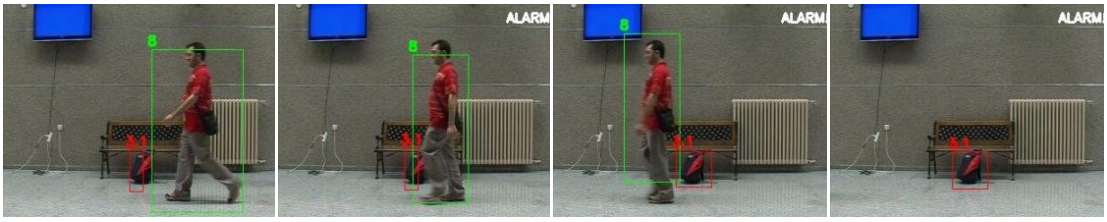
Integration of the abandoned object detection into the tracking system may allow the person who left luggage unattended to be tracked and detected. This method is successful to identify the owner of the abandoned object if a person who left the luggage will stay near it until it is detected as an abandoned object. However, when the luggage left and the owner of the luggage exits from the FOV it would not be possible to find the owner without making some extensions to the system. Therefore, association of living and nonliving objects is essential and necessary as it allows finding the owner of unattended luggage.



Frame #168 Frame #199 Frame #344 Frame #415



Frame #427 Frame #475 Frame #503 Frame #531



Frame #544 Frame #556 Frame #562 Frame #589

Figure 4.8: Abandoned object detection. Person 3 leaves her backpack (object 3.1) on the floor. After it was detected as an abandoned item, temporary occlusions due to moving persons 5 and 6 do not cause the system to fail. The alarm is raised (Frame #556) after the person owning the backpack leaves.

In this thesis, the nonliving object is detected as abandoned object when its owner leaves the FOV and the alarm is set off after *N* frames passed. The alarm is removed immediately when the nonliving object is removed. To prevent false alarms in the case of merging of objects, the object's owner is checked whether it is occluded and formed a new object or not (Figure 4.8).

CHAPTER 5

EXPERIMENTAL RESULTS AND COMPARISONS

In this chapter, the datasets, test and application environment, testing scenarios to evaluate proposed methods in Chapter 3 and Chapter 5 and the experimental results of these methods are presented. The strengths, the weakness, comparisons of these methods with the existing studies are given and the obtained results in different scenarios are illustrated.

5.1. Dataset

To evaluate the performance and to verify the robustness of the method adaptive mean-shift for automated multi object tracking in visible band (presented in Section 3.1), PETS 2006 dataset [51] have been used. This dataset contains sequences taken in a real world public environment and it includes busy scenarios. Video sequences include people walking with their luggage as single or as a part of a larger group and multiple occlusions occur in all scenarios. These videos were captured with digital video (DV) cameras, in phase alternate line (PAL) standard with a 720 x 576 resolution and 25 frames per second and compressed as Joint Photographic Experts Group (JPEG) image format [52].

The proposed method multimodal approach for individual tracking of people and their belongings (Chapter 4) has been tested for 14 different scenarios with various sizes and colors of bags such as black luggage, red handbags and dark blue backpack in crowded scenes and when multiple occlusions exist. To prove the proposed method is not affected from hot objects in the scene, such as heating system,

radiators and so forth, method has also been tested in such environments. As there were no public datasets for abandoned object detection and only a few limited set for object tracking having both modalities, we captured our own dataset for various scenarios. In addition to that, this dataset's thermal band sequences has also been used to test adaptive mean-shift for automated multi object tracking in thermal band which is presented in Section 3.2. Table 5.1 illustrates the details of each image sequences. Video sequences are publicly available at http://ii.metu.edu.tr/content/visible-thermal-tracking.

Table 5.1: Details of the videos used in the evaluation of multimodal approach for individual tracking of people and their belongings.

Scenario	Number of Frames	Number of Living Objects	Number of Non-living Objects	No. of Alarms	General Description of Scenario
Set1	208	1	1	0	No alarm case, backpack is carried
Set2	191	1	1	0	No alarm case, handbag is carried
Set3	539	1	1	0	No alarm case, luggage is carried
Set4	850	3	1	1	Luggage left unattended, multiple occlusion of living objects
Set5	590	>5	1	1	Backpack left unattended, multiple occlusion of living and nonliving objects
Set6	1081	4	1	1	Handbag left unattended, multiple occlusion of living and nonliving objects
Set7	341	2	0	0	No alarm case, occlusion exits
Set8	122	1	1	0	No alarm case, backpack is carried, running living object
Set9	730	1	1	0	No alarm case, backpack is carried, blocking of nonliving object
Set10	450	2	2	0	No alarm case, occlusion of living and nonliving objects
Set11	740	3	1	1	Abandoned luggage, Multiple occlusion of nonliving objects after it is left unattended, Heating system on
Set12	1460	>5	>5	0	No alarm case, multiple occlusion of living and nonliving objects, heating system on
Set13	493	1	1	1	Backpack firstly left unattended and then removed
Set14	409	1	1	1	Luggage firstly left unattended and then removed

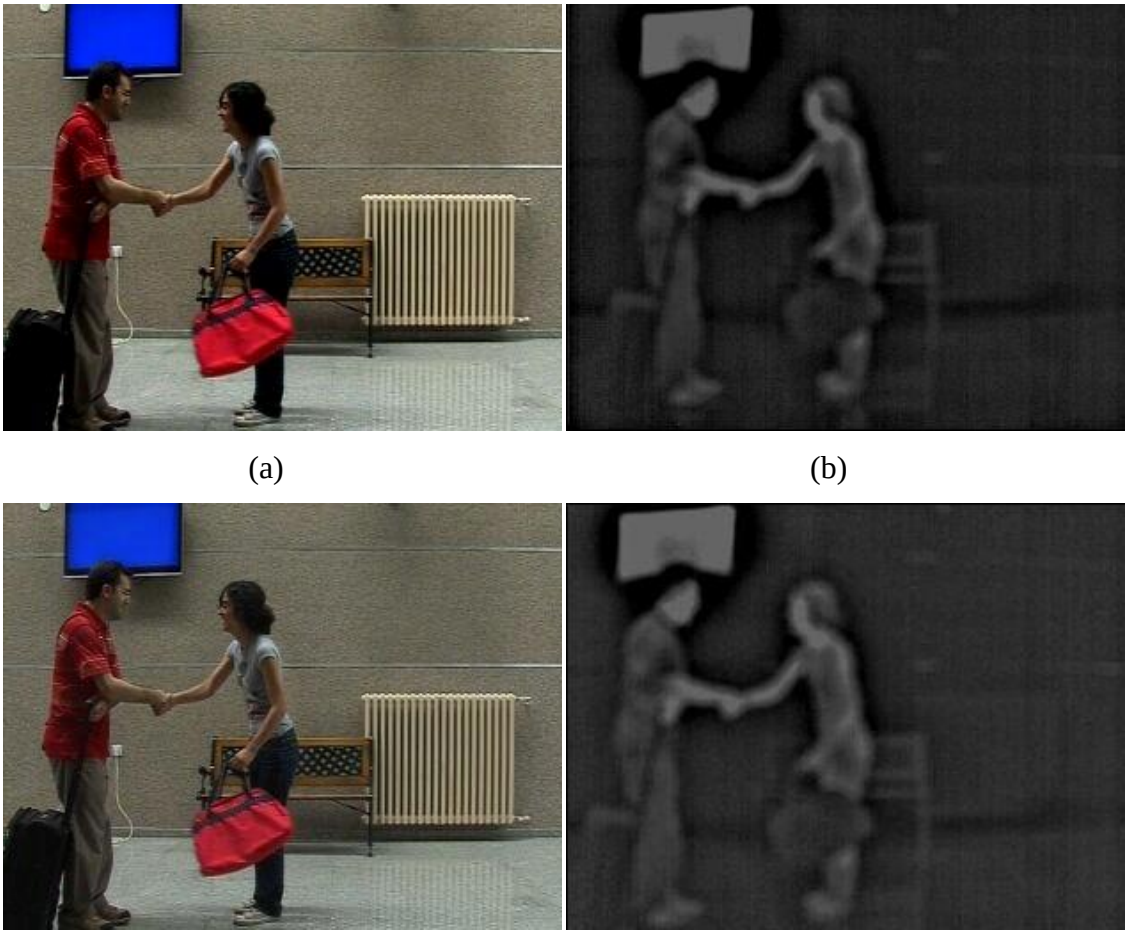
While capturing these scenarios, for thermal video sequence OPGAL EYE-R640 un-cooled infrared camera which captures 25 frames/second at 320x240 resolution and for visible band, Sony HDR-HC1 camera to capture 320x240 images have been used and the scenarios were captured simultaneously.

While capturing, both thermal and visible band cameras were adjusted to capture a similar FOV. However, it was not practically possible to capture exactly the same FOV for both thermal and visible band cameras since these cameras have different parameters such as different sensor types and lenses. Therefore, to track living and nonliving objects and detect abandoned nonliving objects, firstly, images captured from thermal and visible cameras were registered. A crop operation was performed for both thermal and visible band frames to set almost same FOV for both thermal and visible images. Then, homography was performed manually by selecting reference points in both thermal and visible domain for the image registration. To find corresponding pixels of each pixel, homography matrix was constructed. To obtain homography matrix Eq. 5.1, 5.2 and reference points selected from both thermal and visible images were used.

$$V_{ref} = H \times T_{ref} \tag{5.1}$$

$$H = V_{ref} \times T_{ref}^{-1} \tag{5.2}$$

where V_{ref} is the reference point matrix for visible domain, T_{ref} is the reference point matrix for thermal domain, and H is the homography matrix for registration. The more pixels are selected, the better registration results can be obtained. In this study, 20 reference points were selected for each dataset. Once the capture and homography parameters are obtained, these parameters can be used without changing as long as the camera positions are not changed. In this thesis, registrations of the videos are done offline and the results of the registrations are used as an input of the subsequent steps. An example registration result is shown in Figure 5.1.



(a) (b)

(c) (d)

Figure 5.1: An example of image Registration result, (a) Visible image before registration, (b) Thermal image before registration, (c) Visible image after registration, (d) Thermal image after registration

5.2. Test and Application Environment

Figure 5.2: System Setup used for Capturing Video

The system setup consists of thermal and visible camera is shown in Figure 5.2.

The methods proposed in Chapter 3 and 4 have been implemented using C++ with *OpenCV* (Open Computer Vision) library and the registration is implemented in Matlab.

OpenCV is an open source computer vision library written in C and C++ developed by Intel. It could run under Linux, Windows, and Mac OS X. It includes more than 500 functions in different areas of computer vision such as medical imaging, security, user interface, camera calibration and so forth [53].

5.3. Experimental Results

In this section, the test results of the proposed methods in Section 3 and 4, and their comparisons with the existing studies are presented.

5.3.1. Adaptive Mean-Shift for Automated Multi Object Tracking

While applying adaptive mean-shift for automated multi object tracking in visible band (Section 3.1), to perform background subtraction (section 3.1.2) the number of Gaussians were chosen as 4, background model learning rate α was taken as 0.0002, threshold on the squared Mahalanobis distance was taken 16 which means 4 standard deviations in order to provide 99% confidence and initial standard deviation was

taken as 11. To remove shadows (section 3.1.3) α maximum value for the darkening effect of shadows on the background was chosen as 0.6, β the upper bound of the darkening effect was chosen as 0.9, T_s was defined as 0.6 and T_H was used as 0.9. To remove noise (section 3.1.3) the minimum object density not to classify as a noise was chosen as 0.4 and number of pixel threshold was used as 1000. YCrCb color space was preferred as in this color space luminance and chrominance layers are represented separately. For mean shift tracking, three dimensional (3D) (Y, Cr, Cb) histogram was used, histogram bin was taken as 32x32x32, the number of mean-shift iterations to find the new location of trackers in the following frame was taken as one (Since bounding box of object coming from foreground detection is used as a mask, section 3.1.5) and the search area of the mean shift tracker was decreased. Distance threshold while finding the correspondence of object in section 3.1.5 was taken as 25 pixels, and size threshold was defined as 1.3.

Example results of the proposed method are given in Figures 5.3, 5.4, and 5.5. As it is seen from the results, our method successfully detects occlusions, split and finds the correspondence after merging of objects. Multiple occlusions are also handled. By refreshing the trackers, mean shift tracking becomes adaptive to changes in objects' size and shape. By comparing the numbers of objects in consecutive frames, new objects or objects which are leaving the scene are immediately detected. Using foreground detection in initialization step and the update mechanism (section 3.1.5) tracking system becomes fully automatic. Additionally, by removing shadows, accuracy is increased and the false positives are reduced.

Frame # 25

Frame #49

Frame #50

Frame #105

Frame #159

Frame #207

Frame #219

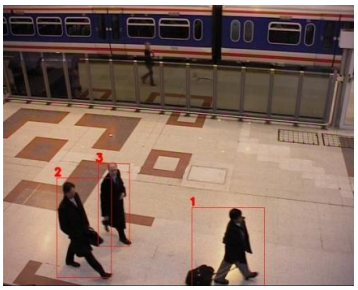
Frame #233

Frame #250

Frame #264

Frame #265

Frame #269



Frame #282

Frame #302

Frame #303

Figure 5.3 (continued)

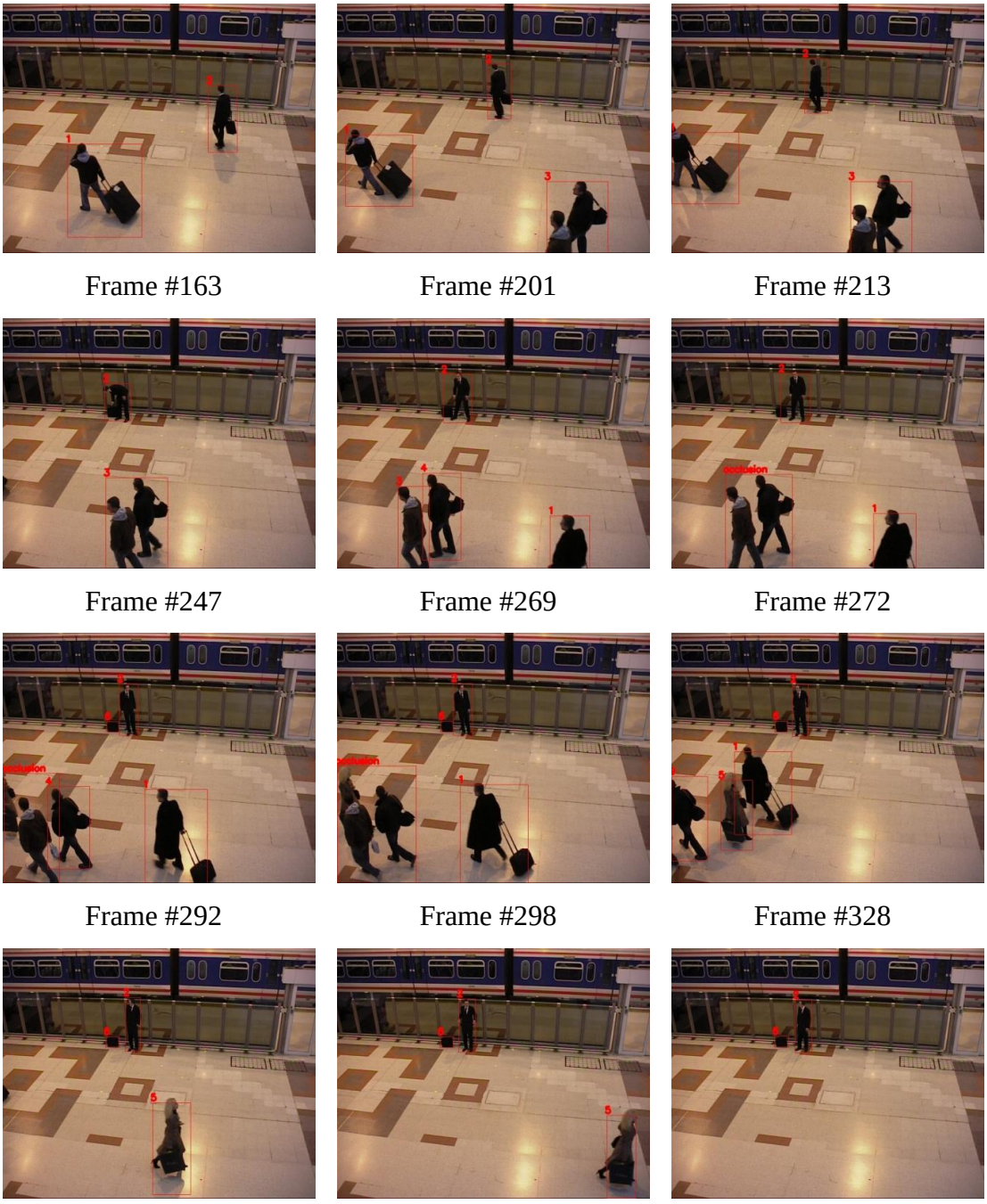
Frame #306 Frame #307 Frame #309

Figure 5.3: An example result for the adaptive mean-shift for automated multi object tracking in visible band (Multiple occlusion handling).

In Figure 5.3, an example result for the adaptive mean-shift for automated multi object tracking in visible band is presented. In this scenario, firstly an object is tracked until the object exits from the FOV and it is marked as object one. Then, another person carrying a luggage comes into the scene and as this method does not contain object discrimination step, the luggage and the living object are extracted as a single object and marked together. There was not any object marked before hence, index one was empty and it is given to these objects. After a while, second and the third objects enter the scene and multiple occlusions are happened. In each occlusion, second object occlude the third object and a new object is formed (marked as occlusion). When the split occurs, the correspondence of second and the third objects are found and they are shown correctly.

Frame #75 Frame #101 Frame #127

Figure 5.4 (continued)



Frame #163 Frame #201 Frame #213

Frame #247 Frame #269 Frame #272

Frame #292 Frame #298 Frame #328

Frame #380 Frame #431 Frame #561

Figure 5.4: An example result for the adaptive mean-shift for automated multi object tracking in visible band (Multiple occlusion handling, group of people exist)

In Figure 5.4, cases; tracking a group of object and multiple occlusions which could happen before the object which is occluded is not defined yet are demonstrated. In this scenario, a group of people marked as object three enter the scene and at Frame

#269 they split, since there is not a predetermined occlusion step, as a result of this splitting, new objects are compared with the objects previously in the scene, no match is found and these two new objects are marked as object three (index three was empty) and four. Then, multiple occlusions happen and the merged objects are marked as occlusion. After splitting they take their old index correctly and the newly entered object which is occluded while entering the FOV marked with an index. Additionally, since we are not aiming to discriminate objects, people's (object two) belongings (object six) are also found as a new object when they split from their owner and they are also tracked until they merge with its owner. Even though object two and six is nearly stationary for more than 427 frames, they could still be tracked correctly.

Frame #26

Frame #72

Frame #87

Frame #91

Frame #135

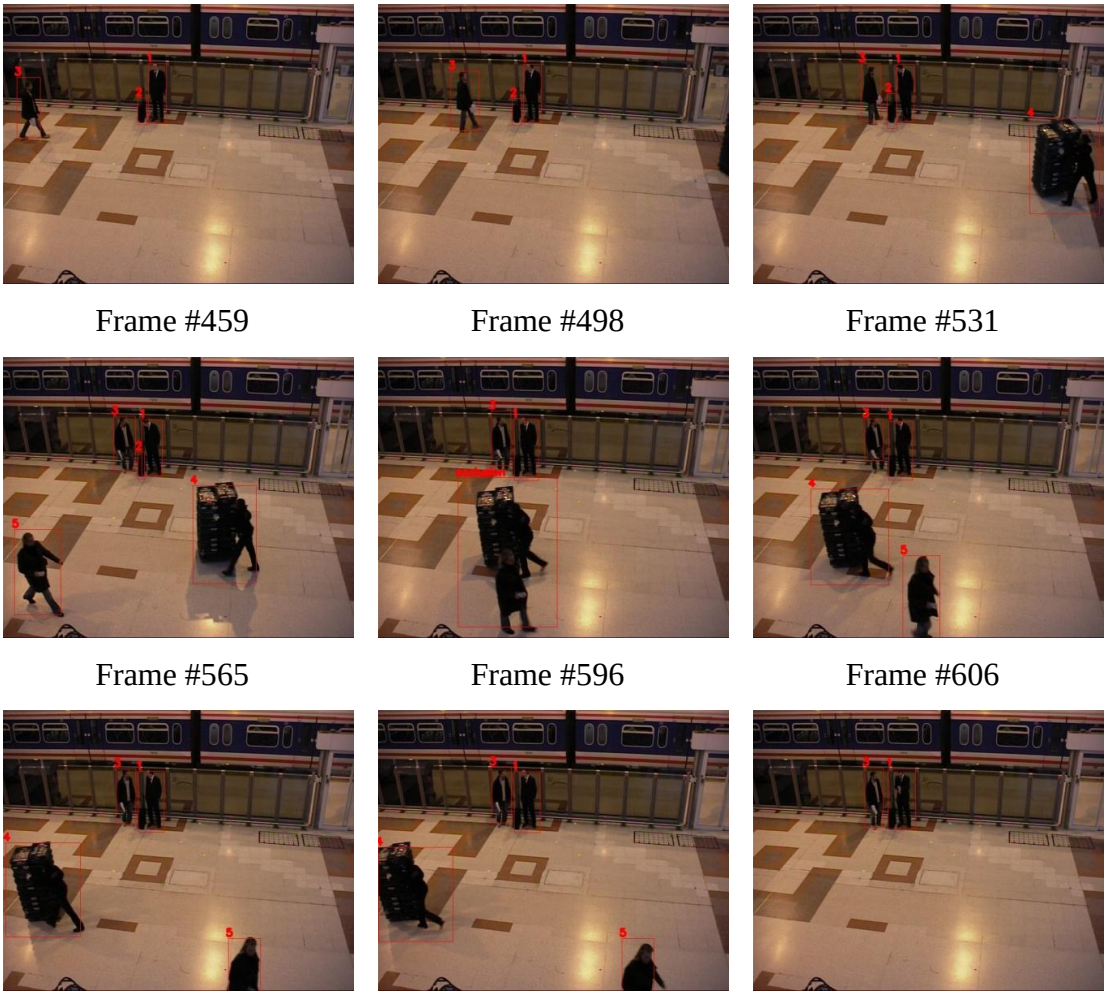
Frame #236

Frame #250

Frame #276

Frame #327

Figure 5.5 (continued)



Frame #459 Frame #498 Frame #531

Frame #565 Frame #596 Frame #606

Frame #625 Frame #630 Frame #671

Figure 5.5: An example result for the adaptive mean-shift for automated multi object tracking in visible band 2 (Occlusion handling, extreme shadow case)

In Figure 5.5, an extreme shadow case is presented. In this scenario, by eliminating shadows with the shadow removal algorithm given in Chapter 3.1.3, possible false positives are averted. Additionally, the bounding box of the object is extracted more correctly which could affect the tracking adversely by increasing the inclusion of background into the tracker. On the other hand, similar to Figure 5.4, the bag is also identified as an object when it splits from the owner and the nearly stationary owner and the fully stationary bag are detected and tracked until the end of the scenario. Occlusion which may cause false tracking is also handled in this case.

Similar results have been observed for all scenarios of PETS 2006.

In addition to the tests done with PETS 2006 dataset, adaptive mean-shift for automated multi object tracking in visible band has also been compared with standard mean shift tracking. While performing standard mean shift tracking two approaches were considered:

- A minimum bounding box including all parts of the objects (external box) is selected. This results in some background information also included in the kernel.
- A box inside of the object is selected (internal box). This is expected to increase the tracking performance as no background information is included in the kernel.

In Table 5.2, the number of objects that are correctly tracked for all methods is given. While executing standard mean shift tracking, trackers are manually initialized for both approaches and the starting and ending frames are selected as given in Table 5.2. Starting frames are chosen as the frames consisting of all the objects to be tracked as the standard mean shift method cannot automatically detect new objects.

As it is seen in Table 5.2, standard mean shift tracking when a external box is chosen correctly tracked only two objects in six scenarios containing a total of 22 objects although the objects were manually initialized. These correctly tracked objects are almost stationary from the beginning to end. However, objects that are running, walking, carrying a luggage or occluded by other objects were not tracked correctly. On the other hand, standard mean shift tracking when an internal box is chosen is more successful compared to external box selection approach. However, this tracking also exhibits degraded performance in the presence of occlusions. On the contrary, proposed method could correctly track all the objects whether running, walking, carrying a luggage or occluded by other objects.

Table 5.2: Comparison of adaptive mean-shift for automated multi object tracking in visible band method and Standard Mean shift considering correctly tracked objects from the beginning to the end

Scenario	Frames	# of objects in the scenario	Standard Mean Shift Tracking		Proposed Method
			External box # of objects that are correctly tracked	Internal box # of objects that are correctly tracked	# of objects that are correctly tracked
S1-T1-C-Video3	[284–430]	4	0	3	4
S2-T3-C-Video3	[554–686]	5	0	5	5
S3-T7-A-Video3	[815–989]	3	0	2	3
S4-T5-A-Video3	[1237–1786]	3	2	2	3
S5-T1-G-Video3	[467–555]	1	0	1	1
S6-T3-H-Video3	[1002–1700]	2	0	0	2
S7-T6-B-Video3	[505–1382]	4	0	4	4

In addition to this comparison, proposed method and standard mean shift tracking (bounding box approach) have been compared in terms of recall and precision evaluation metrics to evaluate the tracking accuracy. For calculating the recall and precision values; the ground truth information which was found in terms of the bounding boxes of objects for each frame and the tracked object's bounding boxes that the proposed tracking system found were utilized. Recall and precision have been calculated using the equations 5.3 and 5.4. While determining these metrics true positive (TP), false positive (FP) and false negative (FN) have been calculated and used as shown in Figure 5.6.

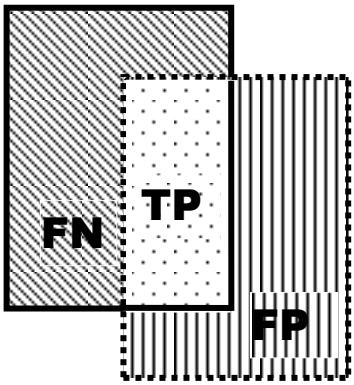


Figure 5.6: Calculation of TP, FP and FNs. Ground truth bounding box for the object is shown in solid lines while the bounding box found by the tracking algorithm is shown in dashed lines.

All TP, FN and FP values are calculated pixel count based where TP is the total number of pixels where the ground truth and tracking system agree on these pixels belonging to an object. FN is the total number of pixels that ground truth denotes the pixel a part of the object while the tracking system cannot detect these pixels as part of an object. FP is defined as the number of pixels that the tracking system finds as an object while ground truth does not agree. In Figure 5.6, dotted area belongs to TP, cross hatching area belongs to FN and vertical hatching area belongs to FP while bounding box drawn in solid lines represents ground truth bounding box and the one drawn in dashed lines is the bounding box found by the tracking system.

$$Recall = \frac{TP}{TP + FN} \tag{5.3}$$

$$Precision = \frac{TP}{TP + FP} \tag{5.4}$$

While calculating recall and precision values in order to compare standard mean shift with proposed method we used scenario *S1-T1-C-Video3* starting from frame number 73 to frame number 191 in order to track a walking object and scenario *S4-T5-A-Video3* starting from 854 to 1004 in order to track a stationary object. Standard mean shift tracking was initialized manually while the proposed method has been run

without any initialization. Illustrations of tracking recall and precision for these two sequences are given in Figures 5.7 and 5.8.

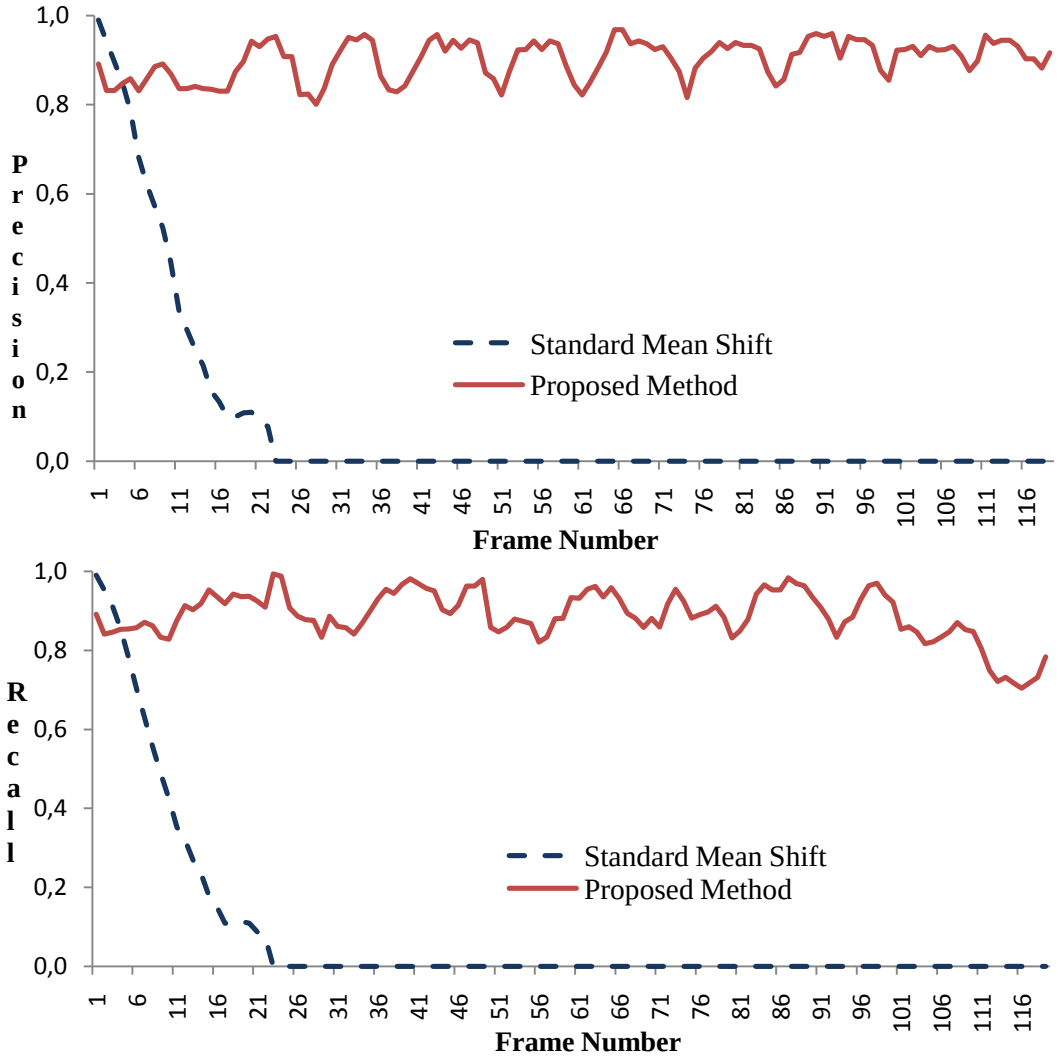


Figure 5.7: Illustration of tracking performance in sequence *S1-T1-C-Video3*, tracking a walking man. Recall and precision are plotted against the frame number at top and bottom respectively.

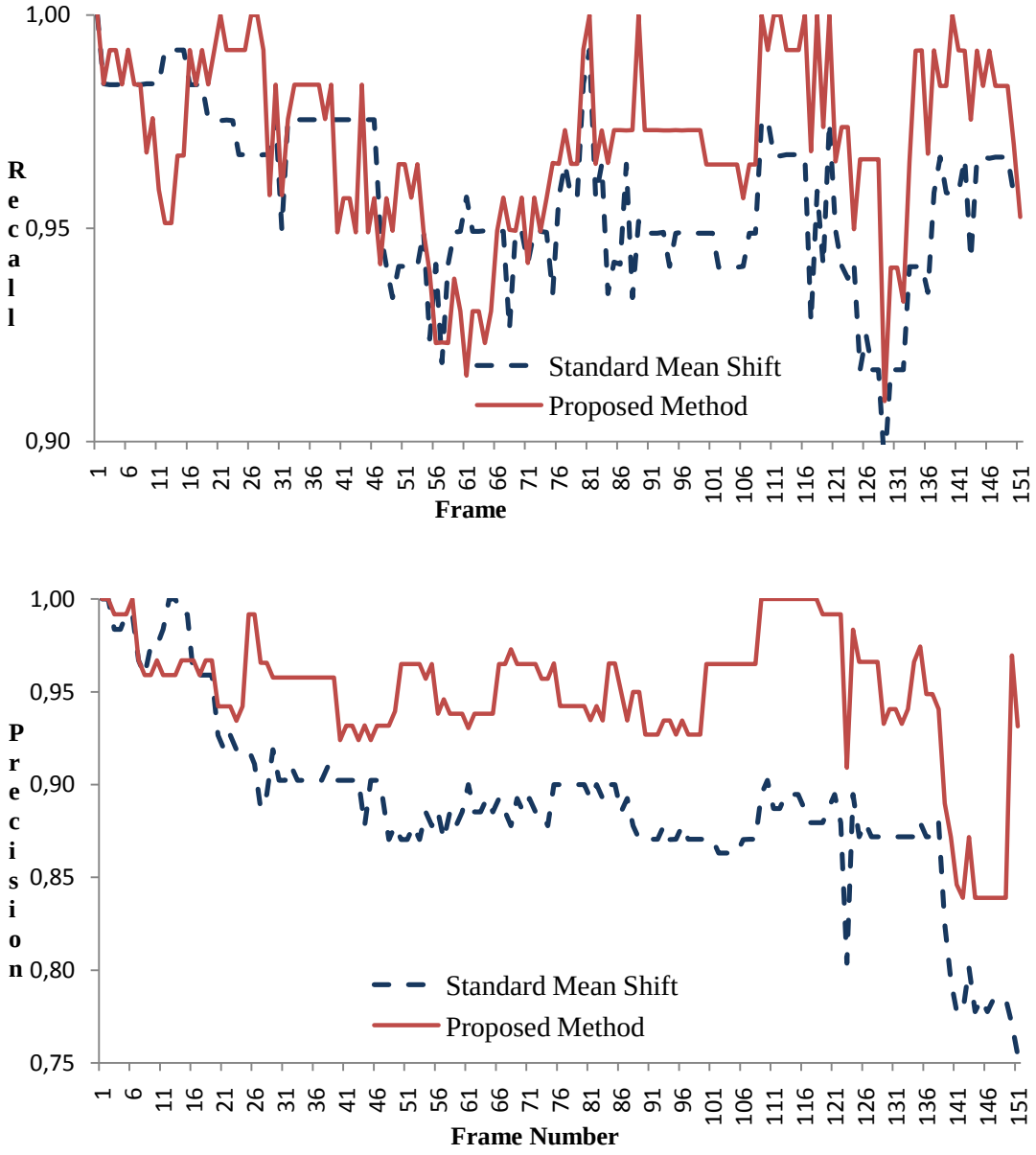


Figure 5.8: Illustration of tracking performance in sequence *S4-T5-A-Video3*, tracking a stationary man. Recall and precision are plotted against the frame number at top and bottom respectively.

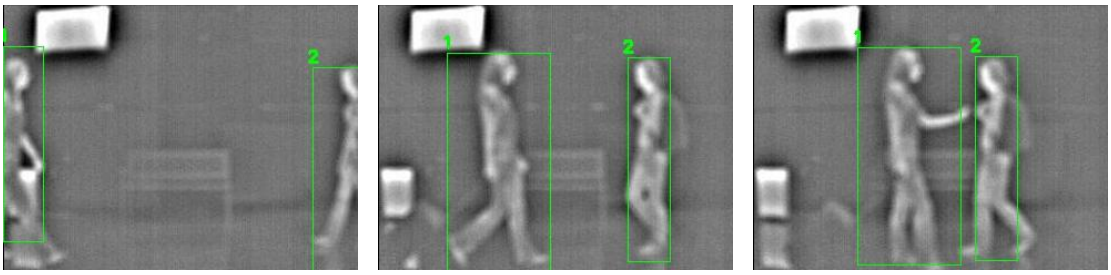
As it is seen from these figures, the proposed method's performance in overall is better than standard mean shift whether object is stationary or moving. For the video sequence with moving object, standard mean shift fails a few frames later, whereas proposed method can track moving object with recall 0.89 and precision 0.9 in average. For the video sequence with the stationary object, although the standard mean shift could track the object, the performance of the proposed method is better

with recall of 0.97 vs. 0.95 and precision of 0.95 vs. 0.88 in average. On the other hand, while the proposed method can handle occlusions, performance of the standard mean shift method degrades significantly in the presence of occlusions.

While applying adaptive mean-shift for automated multi object tracking in thermal band (Section 3.2), the number of Gaussians were chosen as 4, background model learning rate α was taken as 0.0002, threshold on the squared Mahalanobis distance was taken 16 which means 4 standard deviations in order to provide 99% confidence and initial standard deviation was taken as 11 (section 3.2.2) to perform background subtraction and to find foreground objects. To remove noise (section 3.2.2) the minimum object density was chosen as 0.4 and number of pixel threshold was used as 1000. For mean shift tracking, 1D (I: infrared) histogram is used, histogram bin was taken as 32, the number of mean-shift iterations to find the new location of trackers in the following frame was taken as one. Distance threshold while finding the correspondence of object in section 3.2.3 was taken as 25 pixels, and size threshold is defined as 1.3. To evaluate this method, same dataset which was used to evaluate the multimodal approach for individual tracking of people and their belongings method (Chapter 5) has been used.

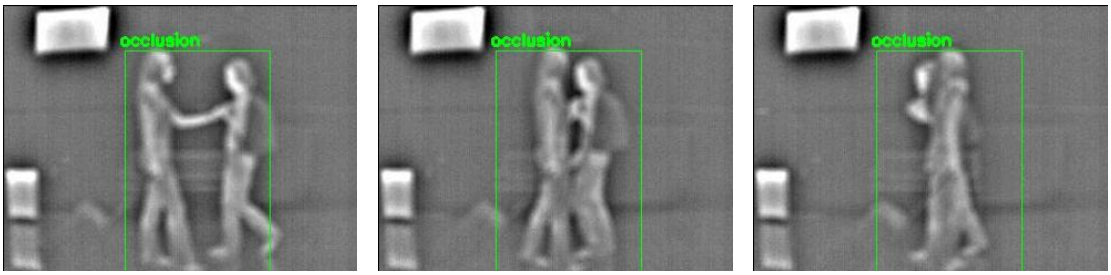
Example results of this method are given in Figures 5.9, and 5.10. According to results, when this method is applied in thermal band, it successfully detects occlusions, split and finds the correspondence after splitting of objects. Additionally, the system become fully automatic, adaptive and new objects or objects which are leaving the scene could immediately be detected. However, since the presented method does not contain a thermal reflection removal method (which is a separate and a difficult task), thermal reflection may cause false positive detections and false positive tracking. To overcome this problem and to reduce or remove the effect of thermal reflections, connected component analysis is used. Moreover, if thermal reflection is connected to object after segmentation step, then it is highly possible that, the bounding box of the object will be larger than the object's real size which could adversely affect the tracking by causing the inclusion of background to foreground object that results in shift or even failure in tracking. However, this shortcoming is also handled using the background modeling as a mask while finding

the next position of the bounding box in the next frame. On the other hand, as it is proposed in the multimodal approach for individual tracking of people and their belongings method (Chapter 5), using visible band in addition to thermal band (fusion of these two different modalities) allows overcoming the thermal reflections.

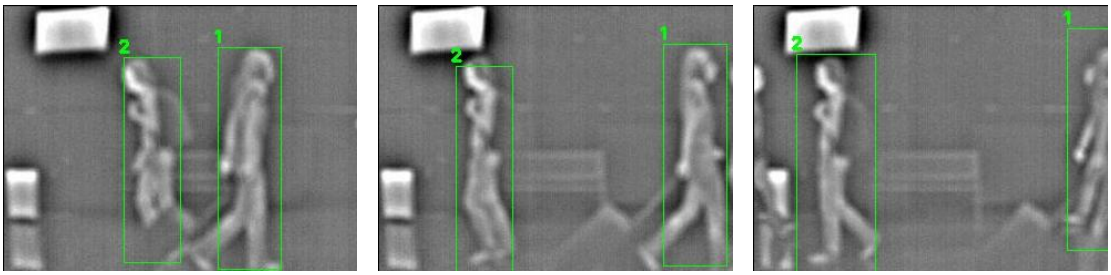


Frame #52 Frame #75 Frame #83

Figure 5.9 (continued)



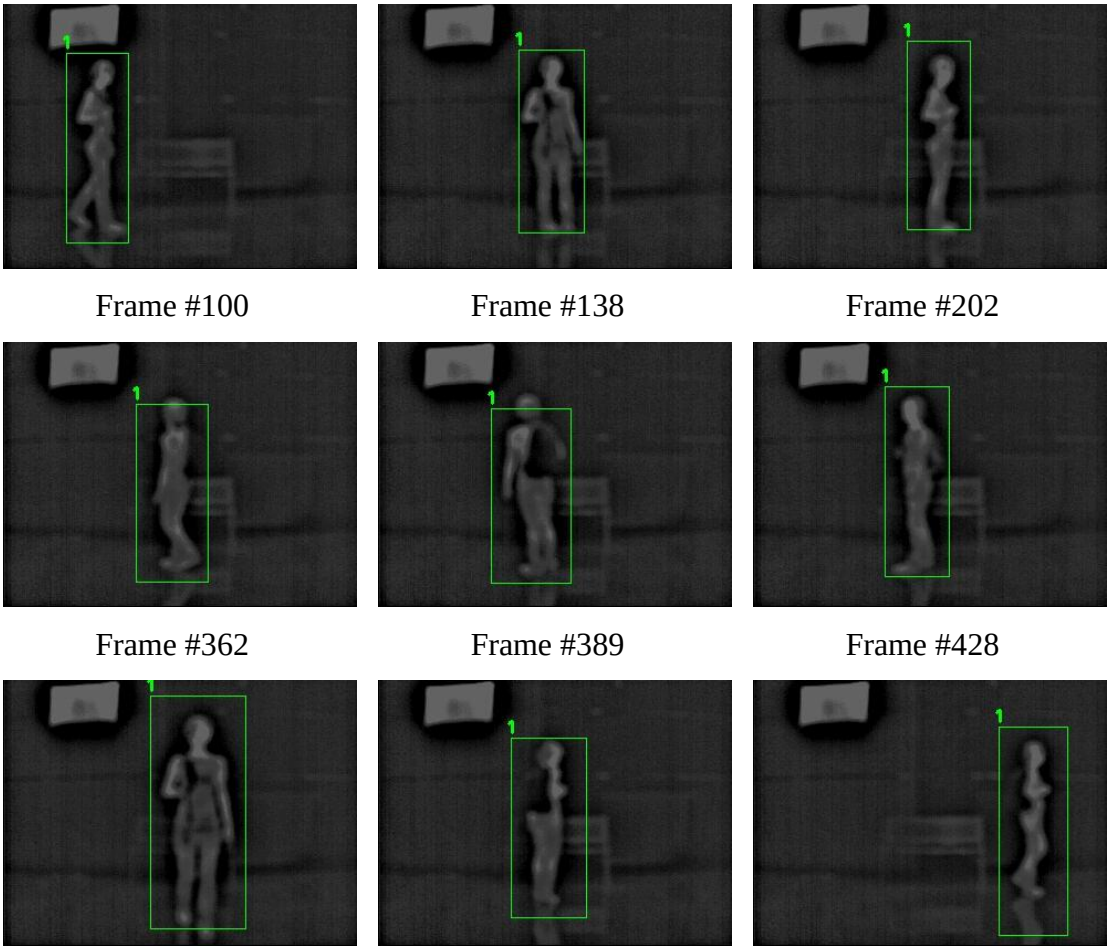
Frame #84 Frame #94 Frame #99



Frame #111 Frame #124 Frame #131

Figure 5.9: An example result for the adaptive mean-shift for automated multi object tracking in thermal band (Occlusion handling)

In Figure 5.9, an occlusion handling is given when adaptive mean-shift for automated multi object tracking in thermal band is applied. As it is seen, the objects are matched correctly after the split. Luggage could not been detected in thermal band, since its pixel value in I band is similar to the scene background.



Frame #100 Frame #138 Frame #202

Frame #362 Frame #389 Frame #428

Frame #532 Frame #648 Frame #704

Figure 5.10: An example result for the adaptive mean-shift for automated multi object tracking in thermal band (Single person tracking with thermal reflection)

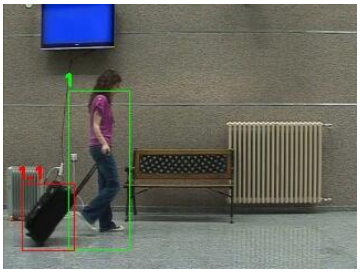
In Figure 5.10 a single person tracking with thermal reflection is shown. As it is seen, the thermal reflection is mostly ignored and does not cause any false tracking in this scenario. Additionally, when the reflection was connected to the object (Frame #704), it has also been tracked as if it is a part of the object, causing a bounding box larger than the object itself.

When adaptive mean-shift for automated multi object tracking in thermal domain and in visible domain are compared, the tracking performance of the method in visible domain is observed to be better than the performance in thermal domain. Firstly, the thermal reflection which does not exist in visible band may cause false objects

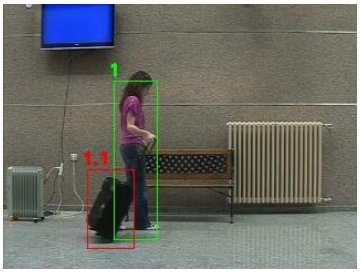
detections and tracking in thermal video. Moreover, for occlusion and split condition (Figure 3.8), using color histogram (used as a feature in visible domain) instead of using size (utilized in thermal domain) feature to find the correspondence of object after splitting occur is more robust as it is possible that object's size may change after split (for instance, object may sit down or get further away from the camera). As a result, the fusion of these two modalities solves the possible problems and a more robust and efficient tracking could be performed.

5.3.2. Multimodal Approach for Individual Tracking of People and Their Belongings and Abandoned Object Detection

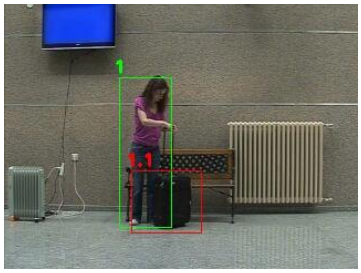
Similar to the method proposed in Section 3, while performing multimodal approach for individual tracking and abandoned object detection, the number of Gaussians is chosen as 4, background model learning rate α is taken as 0.0002, threshold on the squared Mahalanobis distance is taken 16 and initial standard deviation is taken as 11. While applying LIO (Section 4.3) to calculate the MAT I_{max} is taken as 255 and I_{min} is taken as 0 (since the image is normalized to gray-scale image one step before) hence, T is always used as 128. To remove noise, the minimum object density is chosen as 0.4 and number of pixel threshold is used as 1000. For mean shift tracking, 3D (Y, Cr, Cb) histogram is used, histogram bin is taken as 32x32x32, the number of mean-shift iterations to find the new location of trackers in the following frame is taken as one. Distance threshold while finding the correspondence of object is taken as 25 pixels, and size threshold is defined as 1.3. The period of time to alert the system while finding abandoned object (Section 4.8) is chosen as 25 frames (1 sec).



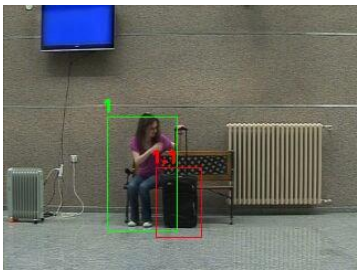
Frame #86



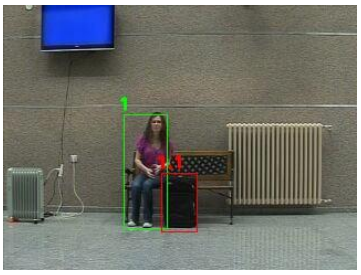
Frame #100



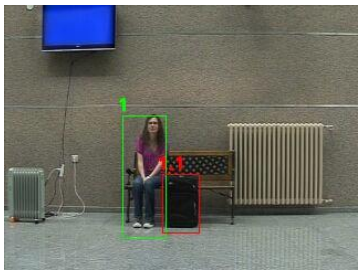
Frame #133



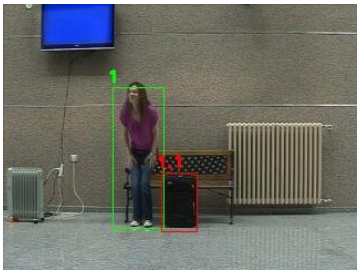
Frame #205



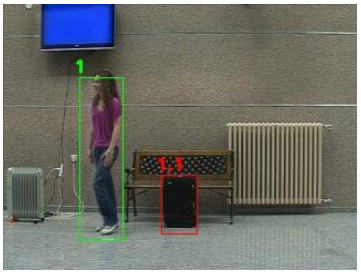
Frame #250



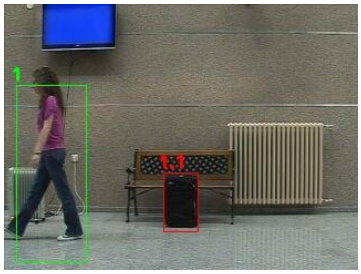
Frame #337



Frame #475



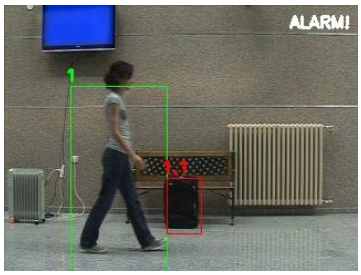
Frame #502



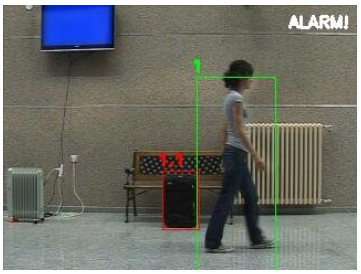
Frame #525

Frame #544

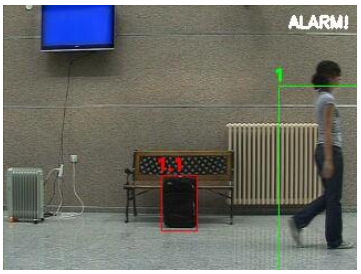
Frame #551



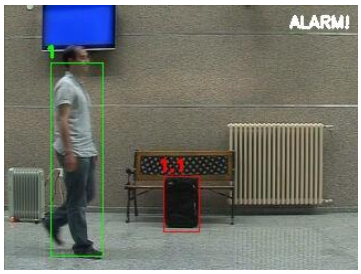
Frame #586



Frame #610

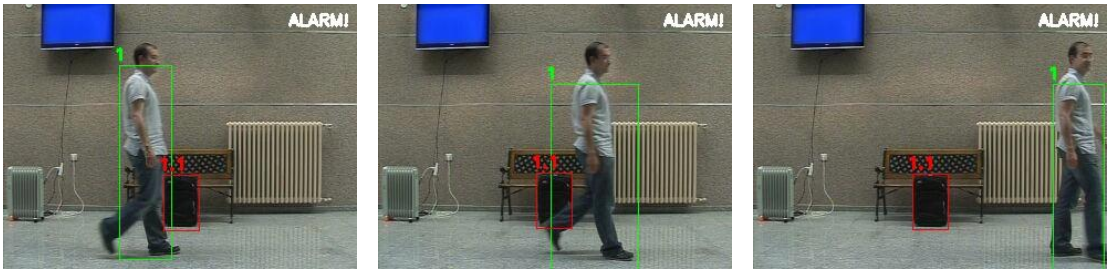


Frame #629



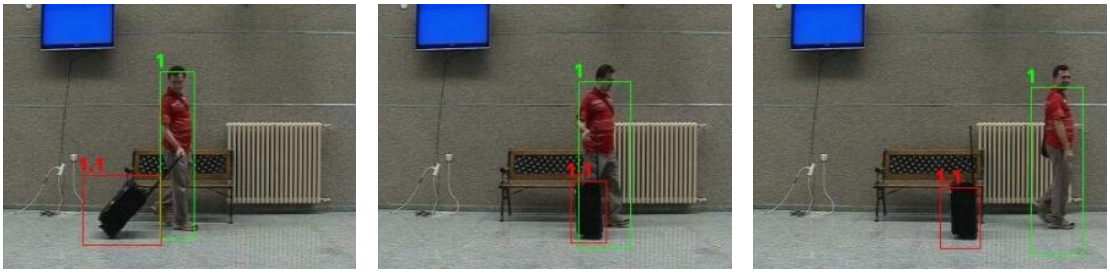
Frame #676

Figure 5.11 (continued)



Frame #689 Frame #703 Frame #724

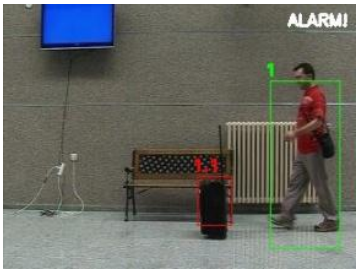
Figure 5.11: An example result for multimodal approach for individual tracking and abandoned object detection, Person 1 brings her luggage (object 1.1) and leaves, which causes an alarm to be set. The luggage is left stationary more than 618 frames. There are multiple occlusions due to other people during this period. Although other people marked as object 1, since there is no object marked as 1, alarm continues since the luggage is not removed.



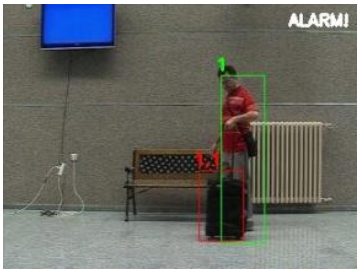
Frame #76 Frame #121 Frame #153

Frame #159 Frame #169 Frame #256

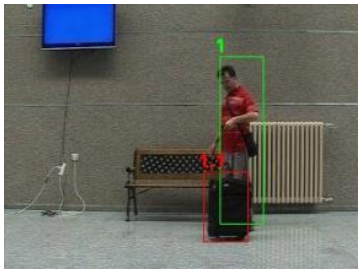
Figure 5.12 (continued)



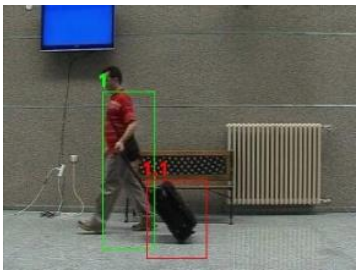
Frame #299



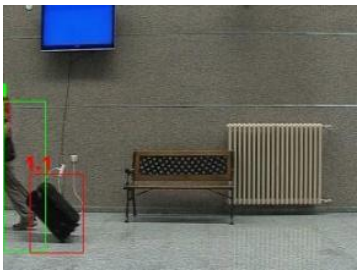
Frame #331



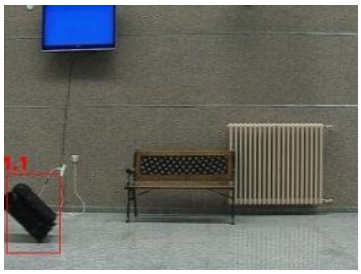
Frame #332



Frame #377

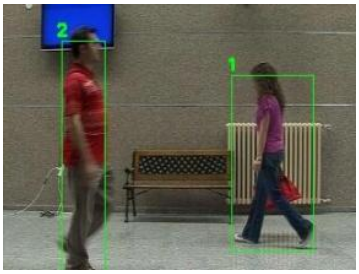


Frame #401

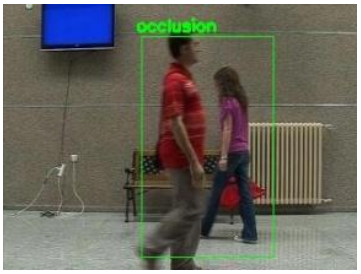


Frame #406

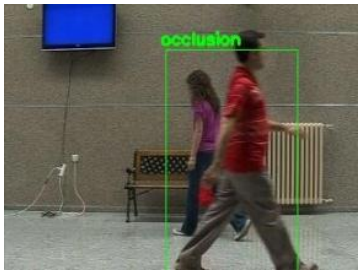
Figure 5.12: An example result for multimodal approach for individual tracking and abandoned object detection, Person 1 brings his luggage (object 1.1) and leaves, which causes an alarm to be set. The luggage is stationary for 229 frames. Then he comes back and takes his luggage, this causes alarm to be removed.



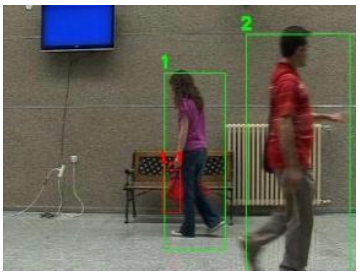
Frame #102



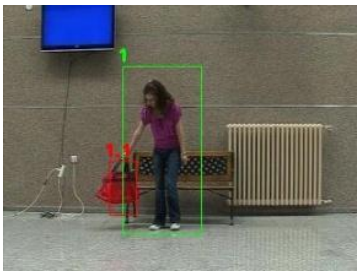
Frame #111



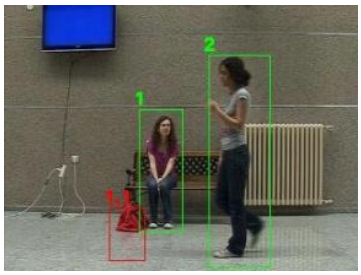
Frame #118



Frame #123

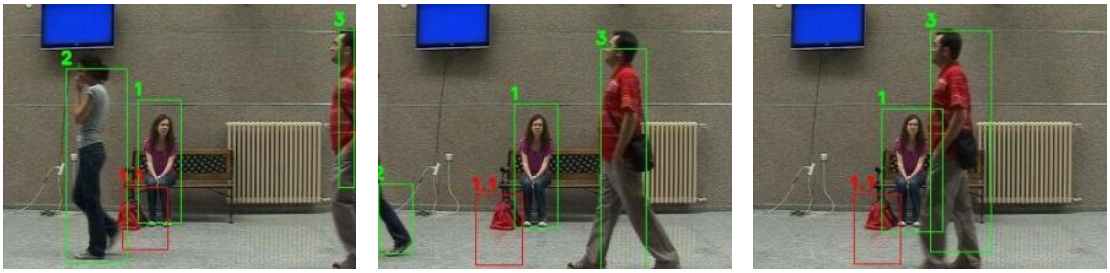


Frame #143

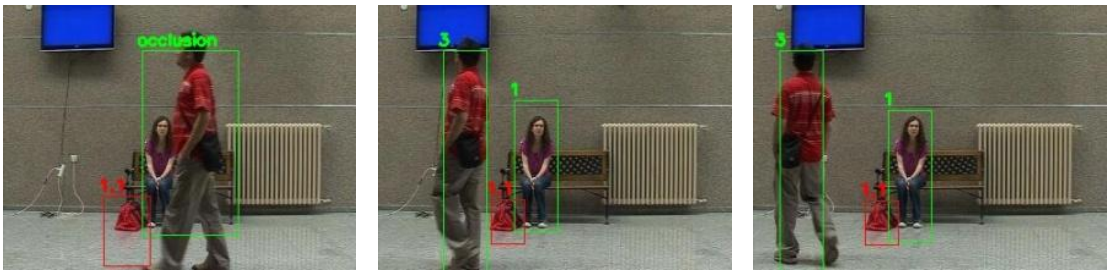


Frame #226

Figure 5.13 (continued)



Frame #253 Frame #275 Frame #285



Frame #286 Frame #312 Frame #323

Figure 5.13: An example result for multimodal approach for individual tracking and abandoned object detection, No false alarm is given although handbag (object 1.1) is stationary for 913 frames as its owner (person 1.1) stays next to it. There are multiple occlusions due to other people during this period.

As it is seen from the example results (Figures 4.8, 5.11, 5.12 and 5.13), the proposed method is applicable for individual tracking of objects such as people and their belongings. Belongings (nonliving objects) are correctly associated to living objects in the presence of occlusion, split and merging of living and nonliving objects and abandoned objects are immediately detected. It also proposes an adaptive, fully automatic mean shift tracking.

To evaluate the tracking performance of multimodal approach for individual tracking the recall and precision values have been calculated using the equations given in Section 5.3.1 and the results are illustrated in Figure 5.14. While calculating the recall and precision values against the frame number the living object in Set1 (which is detailed In Table 5.1) have been used.

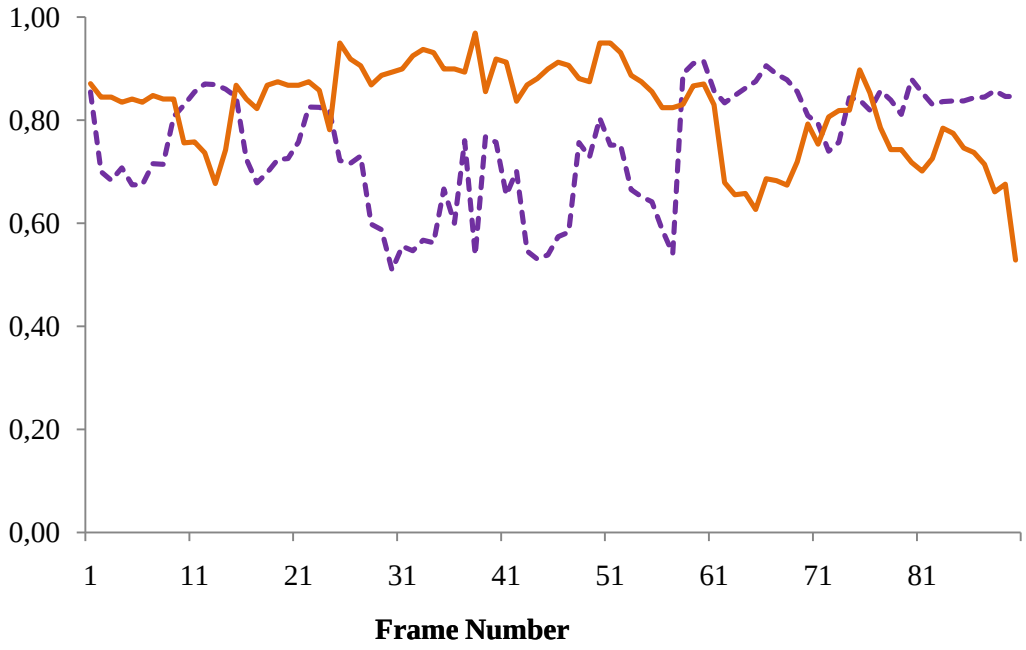


Figure 5.14: Illustration of tracking performance for multimodal approach for individual tracking and abandoned object detection. Recall (purple line) and precision (orange line) are plotted against the frame number.

As it is seen from Figures 5.14, the proposed method (multimodal approach for individual tracking and abandoned object detection) could track moving object with recall 0.75 and precision 0.82 in average. When these results are compared with the results obtained from the method adaptive mean-shift for automated multi object tracking in visible band (Figure 5.7 and 5.8), there is a decrease both in recall and precision values in average. This is because the performance of multimodal approach for individual tracking and abandoned object detection method (Figure 5.14) affected adversely by the performance of the fusion step while detecting the objects. For instance in many scenarios some body parts of the living object (especially hair and feet; hair in general is cooler than body temperature and shoes prevent body heat to be radiated) is removed as a result of fusion step, since it usually could only been detected in visible band while could not be detected in thermal band and eliminated as noise.

On the other hand, multimodal approach for individual tracking and abandoned object detection is also compared with abandoned object detection methods in [46] and [54]. In these studies, a pixel based abandoned object detection algorithm based on long-term and short-term background modeling is proposed. Therefore, these methods do not contain an object detection method and not based on object tracking. In [46], only visible band is used while thermal and visible band are fused in [54].

Table 5.3: Performance of proposed method and methods [46] and [54]

	Robust abandoned object detection using dual foregrounds [46]		**Abandoned object detection using thermal and visible band image fusion [54]**		**Multimodal approach for individual tracking and abandoned object detection**	
Scenario	**True Alarm Detection**	**# of False Alarms**	**True Alarm Detection**	**# of False Alarms**	**True Alarm Detection**	**# of False Alarms**
Set1	NA	**0**	NA	**0**	NA	**0**
Set2	NA	**0**	NA	**0**	NA	**0**
Set3	NA	**0**	NA	**0**	NA	**0**
Set4	**1**	1	**1**	1	**1**	**0**
Set5	**1**	1	**1**	1	**1**	**0**
Set6	**1**	2	**1**	1	**1**	**0**
Set7	NA	**0**	NA	**0**	NA	**0**
Set8	NA	**0**	NA	**0**	NA	**0**
Set9	NA	**0**	NA	**0**	NA	**0**
Set10	NA	**0**	NA	**0**	NA	**0**
Set11	**1**	1	**1**	1	**1**	**0**
Set12	NA	2	NA	**0**	NA	**0**
Set13	**1**	**0**	**1**	**0**	**1**	**0**
Set14	**1**	**0**	**1**	**0**	**1**	**0**

The false and true detection performances (using datasets listed in Table 5.2) of the proposed method and methods in [46, 54] are given in Table 5.3. In this table, the best results for each scenario are given in boldface and no alarm cases are shown as NA as true detection does not apply to these scenarios. As seen, none of the methods missed true alarm cases. No false alarm was given by the proposed method, while method [46] caused false alarms for Sets 4, 5, 6, 11 and 12 since it detect stationary living objects as abandoned objects and method [54] caused false alarms for Sets 4, 5, 6 and 11 as it generates alarm although the owner of nonliving object stays next to the nonliving object. Both methods [46, 54] lose the abandoned object when the object is stationary for a long time. This is due to these objects become a part of the background in time and they could not detect objects when the contrast between the background and the object is low (background and foreground colors are similar).

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

In this thesis, a set of methods which have been applied only on visible band, only on thermal band and fusion of these two modalities for object tracking purpose were proposed and implemented.

As a first technique, adaptive mean shift for automated multi object tracking in visible data is presented. This method is based on four parts including background subtraction, shadow removal, connected component analysis for noise removal and improved, adaptive mean shift tracking. Improved, adaptive Gaussian mixture model proposed in [6] is utilized for background subtraction and the foreground objects are extracted. Then, a shadow detection scheme which is based on the HSV color space and presented in [12] is used to eliminate shadow which results in increase in the robustness of tracking and reduce in the false positives. To remove noise, connected component analysis was used. Finally, objects are tracked using improved, adaptive mean shift tracking which is a fully automatic multiple objects tracking algorithm based on standard mean shift method and presented in this theses as a new tracking method. In this proposed method, foreground detection is used to make the system fully automatic and the bounding boxes coming from foreground detection step are used as a mask to decrease the search area of the mean shift tracker. In this way, while the tracking accuracy is increased, the required number of iterations to find the new location of object is decreased. Objects entering and leaving the scene are all detected in real-time. By regularly updating the trackers using the foreground information coming result of background subtraction and shadow and noise removal, mean shift became more adaptive to the changes in object size and shape. Occlusion,

split and merging scenarios which are not handled by standard mean shift are also handled without any human intervention.

The proposed method adaptive mean shift for automated multi object tracking in visible data presents an easy to implement, robust and efficient mechanism for automated object tracking in the presence of multiple objects. The evaluation results show that the proposed method is superior to the standard mean shift.

Secondly, adaptive mean shift for automated multi object tracking in thermal band data is proposed. Similar to the application in only visible band, this method consists of background subtraction, connected component analysis for noise removal and improved, adaptive mean shift tracking. In thermal band application, the shadow removal step which was applied in visible band, is not applied. For occlusion and split condition, instead of comparing the color histograms' of merged objects (as applied in visible domain) to matched the objects when a split occur, the sizes' of these objects are compared. Different to the visible domain, thermal reflections which may cause false positive detections and false positive tracking exists in the thermal band. To overcome this problem and eliminate the thermal reflections, connected component analysis is used to remove noise. Moreover, using the background modeling as a mask while finding the next position of the bounding box in the next frame, the shortcoming which may cause the bounding box of the object will be larger than the object's real size which could adversely affect the tracking by causing the inclusion of background to foreground object that results in shift or even fail in the tracker was handled. Furthermore, similar to the results in visible domain this method applied in thermal domain also successfully detected occlusions, split and found the correspondence after splitting of objects, become fully automatic, adaptive and new objects or objects which are leaving the scene could be immediately detected.

When adaptive mean-shift for automated multi object tracking in thermal domain and in visible domain is compared, the tracking performance of the method in visible domain was observed to be better than the performance in thermal domain due to thermal reflection sometimes cause false object detections and tracking in thermal domain and using color histogram instead of object size feature to find the

correspondence of objects after splitting was more robust as applied in visible band. As a result, it is shown that, the fusion of these two modalities solves the possible problems and a more robust and efficient tracking could perform considering these problems.

As a final method, a multimodal approach for individual tracking of people and their belongings and abandoned object detection is proposed. This presented method is a fully automatic multiple object tracking algorithm for indoor environments based on the proposed object tracking methods in visible and thermal bands. In addition to visible band, thermal imaging is used and these two modalities are fused after object detection to allow tracking of people and the objects they carry separately. By using the trajectories of these objects, interactions between them are found. Owners of nonliving objects are determined from the beginning of the scene to the end and abandoned objects can immediately be detected. Better tracking performance is also achieved when compared to using only visible or thermal camera by eliminating the shortcomings of these two modalities. Similar to the method adaptive mean shift for automated multi object tracking in visible or thermal data, the information coming from object discrimination step is used to make the system fully automatic and bounding box information is used as a mask to decrease the search area of the mean shift tracker. By this way, the tracking accuracy and the required number of iterations to find the new location of object is decreased. Using the same method improved, adaptive mean shift tracking for both living and non living objects which were also utilized in adaptive mean shift for automated multi object tracking in visible or thermal data, new objects entering to the FOV and objects that are leaving the scene could all be detected immediately and by updating the trackers, mean shift becomes more adaptable to changes in object size and shape, occlusion, split and merging of objects were handled.

The results show that the method is applicable to real life scenarios. Additionally, it performs favorably with other methods regarding the false alarm rates.

As part of this thesis, we provide a publicly available dataset which may be used for further development and benchmarking. We hope that this would make more

research in this field possible by making –generally difficult to acquire- registered thermal-visible video data available for those who might be interested.

The proposed scheme (a multimodal approach for individual tracking of people and their belongings) could also be used in scenarios other then abandoned object detection where tracking of people and their belongings is beneficial. For example, the scheme could be used to count people with or without bags entering/leaving shops or shopping malls. It could also be used to detect someone picking up or stealing an object by adding proper rules after the proposed tracking system. Furthermore, it could be applied while the living objects are wearing heavy clothing such as coats in order to evaluate whether the segmentation of living objects (LIO step) from thermal band is successful or require additional steps.

As a future work, the proposed methods can be adapted to such new scenarios and tested in an outdoor environment. The algorithms could also be implemented on Graphics Processing Unit (GPU) for installations requiring real-time automatic detection from a high number of cameras.

REFERENCES

[1] Lipton, A. J., Fujiyoshi, H., and Patil, R. S., (1998). Moving target classification and tracking from real-time video. *Proceedings of Workshop Applications of Computer Vision,* 129–136.

[2] Collins, R. T., Lipton, A. J., Kanade, T., Fujiyoshi, H., Duggins, D., Tsin, Y., Tolliver, D., Enomoto, N, Hasegawa, O., Burt, P., Wixson, L. (2000). A System for Video Surveillance and Monitoring: VSAM Final Report, Technical report CMU-RI-TR-00- 12, Carnegie Mellon University.

[3] Piccardi, M., (2004). Background Subtraction Techniques: a review. *IEEE International Conference on Systems, Man and Cybernetics*.

[4] Oliver N.M., Rosario B., and Pentland A. P., (2000). A Bayesian Computer Vision System for Modeling Human Interactions, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 22(8), 831-43.

[5] Stauffer, C. and Grimson, W.E.L. (1999), Adaptive background mixture models for real-time tracking, *Proceedings of IEEE Computer Vision Pattern Recognition*.

[6] Zivkovic, Z. (2004). Improved adaptive Gaussian mixture model for background subtraction, *Proceedings of the 17th International Conference on Pattern Recognition,* 28-31.

[7] Gilbert, A., (2008). Scalable and Adaptable Tracking of Humans in Multiple Camera Systems, PhD Thesis, University of Surrey, UK.

[8] Horprasert, T., Harwood, D., and Davis L.S., (1999). A statistical approach for realtime robust background subtraction and shadow detection. *Proceedings of IEEE Frame Rate Workshop,* 1–19.

[9] McKenna, S.J., Jabri, S., Duric, Z., and Wechsler, H., (2000). Tracking interacting people. *Proceedings of International Conference on Automatic Face and Gesture Recognition,* 348–353.

[10] Conaire, C., Cooke, E., O'Connor, N., Murphy, N., Smeaton, A., (2005). Background Modeling in Infrared and Visible Spectrum Video forPeople Tracking, *Proceedings of IEEE International workshop on Object Tracking and Classification Beyond the Visible Spectrum*.

[11] Porikli, F. and Tuzel, (2003). Human body tracking by adaptive background models and mean-shift analysis, *IEEE Conference on Computer Vision Systems, Workshop on PETS*.

[12] Cucchiara R., Grana C., Piccardi M., and Prati A., (2001). Detecting objects, shadows and ghosts in video streams by exploiting color and motion information, *Proceedings of 11th International Conference on Image Analysis and Processing (ICIAP)*, 360 -365.

[13] Kumar, P., Mittal, A., and Kumar, P., (2008). Study of Robust and Intelligent Surveillance in Visible and Multimodal Framework, *Informatica*, 32, 63-77.

[14] Denman, S., Lamb, T, Fookes, C., Chandran, V. and Sridharan, S, (2010). Multi-spectral fusion for surveillance systems, *Computers and Electrical Engineering*, 36(4), 643-663.

[15] Davis, J. W. and Sharma, V. (2007), Background-subtraction using contour-based fusion of thermal and visible imagery, *Computer Vision Image Understanding*, 162-182.

[16] Torresan, H., Turgeon B., Ibarra-Castanedo, C., Hébert, P., and Maldague, X, (2004). Advanced Surveillance Systems: Combining Video and Thermal Imagery for Pedestrian Detection, *SPIE Thermosense XXVI*, 506–15.

[17] Kumar,P. Mittal, A. and Kumar, P, (2006). Fusion of Thermal Infrared and Visible Spectrum Video for Robust Surveillance, *Proceedings of 5th Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, 528–539.

[18] Conaire, C. Ó, O'Connor, N., Cooke, E., and Smeaton, A, (2006). Multispectral Object Segmentation and Retrieval in Surveillance Video, *Proceedings of International Cpnference of Image Processing (ICIP).*

[19] Conaire, C.O., O'Connor, N.E.O., Cooke, E., and Smeaton, A.F., (2006). Comparison of Fusion Methods for Thermo-Visual Surveillance Tracking, Proceedings of *9th International Conference on Information Fusion* , 1-7.

[20] Krotosky, S.J., Trivedi, M. M., (2008). Person Surveillance Using Visual and Infrared Imagery, *IEEE Transactions on Circuits and Systems for Video Technology*, 18 (8), 1096-1105.

[21] Yilmaz A., Javed, O., and Shah, Mubarek, (2006). Object Tracking: A Survey, *ACM Journal of Computing Surveys*, 38-4.

[22] Dedeoglu Y., (2004). Moving Object Detection, Tracking and Classification for Smart Video Surveillance', Master's thesis, Bilkent University, Department of Computer Engineering, Turkey.

[23] Haritaoglu, I., Harwood, D. and Davis, L.S, (1998). W4: A real time system for detecting and tracking people. *Proceedings of Computer Vision and Pattern Recognition*, 962–967.

[24] Ju, S., Black, M., and Yaccob, Y., (1996). Cardboard people: a parameterized model of articulated image motion. *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, 38–44.

[25] Wren, C. R., Azarbayejani, A., Darrell, T. J., and Pentland, A. P., (1997). Pfinder: Real-time tracking of the human body. *IEEE Pattern Recognition and Machine Intelligence*, 19 (7), 780–785.

[26] Isard, M., MacCormick, J., (2001). Brammble: A Bayesian Multiple Blob Tracker. *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition,* 34 (41).

[27] Bose, B., Wang, X., Grimson, E., (2006). Detecting and Tracking Multiple Interacting Objects Without Class-Specifc Models. Technical report MIT-CSAIL-TR-2006-027 Massachusetts Institute of Technology.

[28] Comaniciu D., Ramesh V., and Meer P., (2003). Kernel-based object tracking, *IEEE Trans. Pattern Analysis and Machine Intelligence,* 25(5), 564–77.

[29] Birchfield S. T., and Rangarajan S., (2005). Spatiograms versus histograms for region-based tracking, *Proceedings of IEEE Conferences on Computer Vision and Pattern Recognition (CVPR),* 1158-63, San Diego, CA USA.

[30] Boonsin M., Wettayaprasit W., and Preechaveerakul L., (2010). Improving of Mean Shift Tracking Algorithm Using Adaptive Candidate Model, *Proceedings of ECTI-CON,* 894 – 98, Chiang Mai, Thailand.

[31] Collins R. T., (2003). Mean-shift blob tracking through scale space, *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR),* 234–40, Wisconsin, USA.

[32] Qifeng Q., Zhang D., and Peng Y., (2007). An adaptive selection of the scale and orientation in kernel based tracking, *Proceedings of IEEE Conf. on Signal-Image Technologies and Internet-Based Systems (SITIS),* 659–64, Shanghai, China.

[33] Parameswaran V., Ramesh V., and Zoghlami I., (2006). Tunable kernels for tracking, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2179-86, New York, USA.

[34] Quast K., and Kaup A., (2009). Scale and shape adaptive mean shift object tracking in video sequences, Proceedings of 17th European Signal Processing Conferences (EUSIPCO), 1513–17, Glasgow, Scotland.

[35] Jiang Z., and Li S., Gao D., (2007). An Adaptive Mean Shift Tracking Method Using Multiscale Images, *Proceedings of the 2007 International Conference on Wavelet Analysis and Pattern Recognition,* Beijing, China.

[36] Porikli F., and Tuzel O., (2005). Multi-kernel object tracking, *Proceedings of IEEE International Conference on Multimedia and Expo (ICME),* 1234-37, Amsterdam, Netherlands.

[37] Ferrando S., Gera G., Massa M., and Regazzoni, C., (2006). A new method for real time abandoned object detection and owner tracking, *Proceedings of ICIP*, 3329–3332.

[38] Auvinet, E., Grossmann, E., Rougier, C., Dahmane, M., and Meunier, J., (2006). Left-luggage detection using homographies and simple heuristics, *in PETS*, 51–58.

[39] Rincn J. M., Herrero-Jaraba J. E., Gmez, J. R. and Orrite-Uruuela, C., (2006). Automatic left luggage detection and tracking using multi-camera, *in PETS*, pp. 59–66.

[40] Krahnstoever P. T. N., Sebastian T., Perera A., and Collins R., (2006). Multi-view detection and tracking of travelers and luggage in mass transit environments, *in PETS*, 67–74.

[41] Smith K., Quelhas P., and Gatica-Perez D., (2006). Detecting abandoned luggage items in a public space, *in PETS*, 75–82.

[42] Guler S. and Farrow M. K., (2006). Abandoned object detection in crowded places, *in PETS*, 99–106.

[43] Miezianko R. and Pokrajac D., (2008). Detecting and recognizing abandoned objects in crowded environments, *Computer Vision Systems*, 241–250.

[44] Denman S., Sridharan S., and Chandran V., (2007). Abandoned object detection using multi-layer motion detection, *Proceeding of International Conference on Signal Processing and Communication Systems.*

[45] SanMiguel Juan C., Martínez José M., (2008). Robust unattended and stolen object detection by fusing simple algorithms, *Proceedings of the 2008 IEEE International Conference on Advanced Video and Signal based Surveillance, AVSS'2008,* 18-25, Santa Fé (NM, USA).

[46] Porikli F., Ivanov Y., Haga T., (2008). Robust abandoned object detection using dual foregrounds, *EURASIP Journal on Advances in Signal Processing*.

[47] Haralick R. M., Shapiro L. G., (1992). Computer and Robot Vision, 1, Addison-Wesley.

[48] Glassner A., (2001). “Fill Er Up!”, IEEE Computer Graphics and Applications, 21, 78-85.

[49] Rosenfeld A., Pfalz J. L., (1966). Sequential Operations in Digital Picture Processing, Journal of the ACM, 13(4), 471-494.

[50] Heriansyah R. and Abu-Bakar S.A.R., (2009). Defect detection in thermal image for nondestructive evaluation of petrochemical equipments, *NDT & E International*, 42(8), 729-740.

[51] Performance Evaluation of Tracking and Surveillance (PETS) 2006 Benchmark Data, http://pets2006.net/, accessed June 2010.

[52] Thirde D., Li L., and Ferryman J., (2006). Overview of the PETS2006 Challenge, *Proceedings of Ninth IEEE Internaitonal Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, 47–50, New York, USA.

[53] Gary, Bradski and Kaehler, Adrian., (2008). *Learning OpenCV*. O'Reilly, 978-0-596-51613-0.

[54] Yigit A., Temizel A., (2010). Abandoned Object Detection Using Thermal and Visible Band Image Fusion, *IEEE Signal Processing, Communication and Applications Conference (SIU)*, Diyarbakır, Turkey.