HUMAN BODY PART DETECTION AND MULTI-HUMAN TRACKING IN SURVEILLANCE VIDEOS

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

BY

HASAN HÜSEYİN TOPÇU

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

MAY 2012

Approval of the thesis:

HUMAN BODY PART DETECTION AND MULTI-HUMAN TRACKING IN SURVEILLANCE VIDEOS

submitted by **HASAN HÜSEYİN TOPÇU** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Adnan Yazıcı
Head of Department, **Computer Engineering** _____

Prof. Dr. Nihan Kesim Çiçekli
Supervisor, **Computer Engineering Department, METU** _____

Assist. Prof. İlkay Ulusoy
Co-supervisor, **Electrical and Electronics Eng. Dept., METU** _____

Examining Committee Members:

Assist. Prof. Sinan Kalkan
Computer Engineering Dept., METU _____

Prof. Dr. Nihan Kesim Çiçekli
Computer Engineering Dept., METU _____

Assist. Prof. Murat Manguoğlu
Computer Engineering Dept., METU _____

Assist. Prof. Alptekin Temizel
Informatics Institute, METU _____

Sezen Erdem, M.Sc.
SST, ASELSAN A.Ş. _____

Date: _____

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

Name, Last Name: HASAN HÜSEYİN TOPÇU

Signature :

ABSTRACT

HUMAN BODY PART DETECTION AND MULTI-HUMAN TRACKING IN SURVEILLANCE VIDEOS

Topçu, Hasan Hüseyin

M.Sc., Department of Computer Engineering

Supervisor : Prof. Dr. Nihan Kesim Çiçekli

Co-Supervisor : Assist. Prof. İlkay Ulusoy

May 2012, 59 pages

With the recent developments in Computer Vision and Pattern Recognition, surveillance applications are equipped with the capabilities of event/activity understanding and interpretation which usually require recognizing humans in real world scenes. Real world scenes such as airports, streets and train stations are complex because they involve many people, complicated occlusions and cluttered backgrounds. Although complex real world scenes exist, human detectors have the capability to locate pedestrians accurately even in complex scenes and visual trackers have the capability to track targets in cluttered environments. The integration of visual object detection and tracking, which are the fundamental features of available surveillance applications, is one of the solutions for multi-human tracking problem in crowded scenes which is studied in this thesis.

In this thesis, human body part detectors, which are capable of detecting human heads and human upper body parts, are trained with Support Vector Machines (SVM) by using Histogram of Oriented Gradients (HOG), which is one of the state-of-the-art descriptor for human detection. The training process is elaborated by investigating the effects of the parameters of the HOG descriptor. The human heads and upper body parts are searched in the region of

interests (ROI) computed by detecting motion. In addition, these human body part detectors are integrated with a multi-human tracker which solves the data association problem with the Multi Scan Markov Chain Monte Carlo Data Association (MCMCDA) algorithm. Associated measurements of human upper body part locations are used for state correction for each track. State estimation is done through Kalman Filter. The performance of detectors are evaluated using MIT Pedestrian dataset and INRIA Human dataset.

Keywords: Human Head Detection, Human Upper Body Detection, Multi-Human Tracking

ÖZ

GÖZETLEME VİDEOLARINDA İNSAN VÜCUT PARÇASI BULMA VE ÇOKLU İNSAN TAKİBİ

Topçu, Hasan Hüseyin
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi : Prof. Dr. Nihan Kesim Çiçekli
Ortak Tez Yöneticisi : Yrd. Doç. Dr. İlkay Ulusoy

Mayıs 2012, 59 sayfa

Bilgisayarlı görme ve örüntü tanıma alanlarındaki son gelişmeler ile gözetleme uygulamaları insanın sahnede tanınmasını da gerektiren olay/aktivite anlama ve yorumlama yetenekleriyle donatılmaya başlandı. Havaalanları, caddeler ve tren istasyonları gibi gerçek dünya sahneleri içerisinde çok fazla insan bulundurması ve insanların örtüşmesi sebebiyle karmaşıktır. Gerçek dünya sahnelerinin karmaşıklığına rağmen, insan bulucular karmaşık sahnelerde bile doğru şekilde insanları konumlandırabiliyor, görsel takip ediciler gürültülü ortamlarda hedefler izleyebiliyorlar. Bu tezde de üzerinde durulan, gözetleme uygulamalarının temelini oluşturan görsel nesne bulma ve izleyicilerin birleştirilmesi kalabalık sahnelerde çoklu insan takibi için kullanılan çözümlerden birisidir.

Bu tezde, görsel insan başı ve vücut üst kısmı tanıma yeteneğine sahip olan insan vücut parçası bulma detektörleri Destek Vektör Makinesi (DVM) yardımıyla eğitilmiştir. Özellik tanımlayıcı olarak, insan tanınmasında en başarılı özellik tanımlayıcılarından biri olan HOG kullanılmıştır. Eğitim süreci HOG parametrelerinin etkisi açıklanarak detaylandırılmıştır. İnsan başları ve üst vücutları hareketi tespiti yardımıyla bulunan alanlarda aranır. Bunun dışında, vücut parçası bulma detektörü, veri eşleştirme problemini Markov Zinciri Monte

Carlo Veri Eşleştirme algoritmasıyla çözen çoklu-insan takip edici ile entegre edilmiştir. Eşleştirilmiş insan üst vücut konum ölçümleri her bir iz için durum düzeltmesinde kullanılmıştır. Durum tahmini Kalman Filtresi ile yapılmıştır. Detektörlerin performansı MIT yaya veriseti ve INRIA insan veriseti kullanılarak değerlendirilmiştir.

Anahtar Kelimeler: İnsan Başı Bulma, İnsan Üst Parçası Bulma, Çoklu İnsan Takibi

To my family and my friends

ACKNOWLEDGMENTS

I would like to thank to Nihan K. Çiçekli and İlkay Ulusoy for their supervision and guidance through the development of this thesis.

I would like to thank to my family and my supportive friends for their belief in me.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

1.1 Motivation

Computer Vision is one of the fields that has an impressive impact on the daily life and military field as well. Streets, airports, train stations, robots and satellites are equipped with various types of cameras. According to [1], the number of cameras in England is approximately 4 millions in 2005. While some of these cameras have the capability of capturing high definition image from distances, some others can capture thermal images or interpret what they see called *Smart Cameras*[2] defining the ultimate goal of the smart cameras as mimicing the human eyes and brain. This challenging purpose has attracted significant interest from academic community as well as the industry and governments. As an example, The Defense Advanced Research Projects Agency (DARPA) has initiated a program called *Mind's Eye*[3] which addresses the problem of absence of visual intelligence capability in unmanned systems.

Visual Surveillance is the most prominent field of Computer Vision. This field has application areas such as traffic monitoring, homeland security, automotive safety and monitoring a scene for detecting abnormal events. Object detection, recognition and tracking are the fundamental features of surveillance systems. Surveillance systems in daily life generally include vehicles and humans. These systems are listed below[4]:

- Access control in special areas
- Person specific identification in certain scenes
- Crowd flux statistics and congestion analysis

- Anomaly detection and alarming

- Interactive surveillance with multiple cameras

The focus of this thesis is human detection and tracking of multiple humans in surveillance videos due to their popularity in academic community and industry. Intelligent visual surveillance provides automated object recognition, tracking of objects, event/activity understanding and indexing/retrieval of visual events compared to conventional visual surveillance [5]. The main goal of this thesis has been producing some beneficial output to be used in an event/activity understanding system. Another motivation is to assist semantic annotation of videos with a robust object/human detection tool.

1.2 Scope of the Thesis

In this thesis, it is aimed to develop a system that has the capability of detecting humans and tracking multiple humans in surveillance videos without a real time requirement. Activity recognition, which is defined as a "complex sequence of actions performed by several humans" [6], requires discriminating the occluded humans in the scene. Therefore a robust human detection mechanism is one of the requirements of this thesis. Although the human detection has been studied for years, it is still a challenging problem due to appearance variations, shape variations and viewpoint variations. The *variability* is regarded as a curse of the Computer Vision by Freeman [7]. Addressing the *variability* problem is within the scope of this thesis.

Visual surveillance applications should be able to track multiple objects, handling the enterance of objects into the scene and exitting of objects from the scene. Therefore another requirement of the thesis is achiving the multi-human tracking for unknown number of humans.

As a result, the scope of this thesis is defined as detecting humans in the scene handling appearance variations and tracking multiple humans handling the dynamic nature of the scene such as appeareance and disapperance of humans.

1.3 Our Approach

In this thesis, the integration [8, 9] of human detection and tracking is adopted for the multi-human tracking problem. People detectors can locate people even in complex scenes [8, 10, 11, 12, 13]. Tracking methods have the ability to find a particular human in image sequences [14], but are severely suffered from crowded scenes [8]. Combining both advantages of human detection and tracking in a single framework is regarded as a promising research direction by Schiele et al [8, 9]. The motion information, which is used in general object tracking problems [14], is used only for reducing the computational overhead of human detection in a scene.

Our overall system has the following parts:

- Firstly, the region of interests (ROIs) that can have potential targets are computed using the motion information.
- Secondly, the human head locations and human upper body locations are detected by the trained detector in the ROIs.
- Thirdly, the detection locations (called *measurements*) are associated to current tracks by the MCMCDA algorithm.
- Finally, the state of each track are updated (corrected) by using associated measurements with Kalman Filter algorithm.

1.4 Contributions

The main contributions of this thesis are as follows:

- The human head and upper body training procedure is examined in detail. The effects of the parameters of the HOG descriptor are investigated for human head training and human upper body training.
- Multi-Scan version of the MCMCDA algorithm [15] is applied to a Computer Vision problem which has already presented for general tracking problems in [15].
- Instead of traditional solutions to object tracking, integration of human detection and multi-human tracking is used [8, 9] which is best applicable to crowded scenes.

1.5 Thesis Organization

This thesis is organized as follows:

Chapter 2 presents related work and gives background methods for human detection, human tracking and data association techniques for multi-human tracking.

Chapter 3, explains the pre-processing steps for detecting the regions that have potential targets.

In Chapter 4, the training steps for visual human head detection and human upper body detection are presented. The effects of the HOG descriptor parameters are examined in detail in this chapter. Also, the results are given in the rest of the Chapter 4.

In Chapter 5, Multi-Scan MCMCDA (Markov Chain Monte Carlo Data Association) algoritm is presented as a data association technique. The state-space model used for multi-human tracking and Kalman Filter algorithm are introduced.

The summary of the thesis and future work are explained in Chapter 6.

CHAPTER 2

RELATED WORK

The detection and tracking of people are the main tasks in visual surveillance [5, 25] and high level human motion analysis is highly dependent on the accuracy of these main tasks [25]. Visual object detection and tracking have been both active research topics for many years due to the growing needs of the society. Surveillance cameras spread over the streets, airports, train stations, etc. in order to monitor the actions and activities of humans, vehicles, and groups of people. While the surveillance equipments are increasing, the need for automated human detection and tracking is being inevitable for the surveillance applications.

There is an extensive literature on the object detection and tracking. In spite of this immense literature, an advanced surveillance application may not still detect a human or track a human for a long period especially in an uncontrolled environment. The complexities coming from the nature of the problem are listed below [7, 14]:

- Nonrigid or articulated objects
- Scene illumination variations
- Partial and full object occlusions
- Complex object motion
- View Variations

Freeman exemplifies the viewing condition variation for the face recognition problem [7]. Computer Vision systems can detect the frontal view faces under good conditions whereas the recognition rates drop significantly for non-frontal faces. Therefore computer vision solves

these problems by imposing the constraints on the motion, appearance, viewing condition or uses priori information such as the number of objects, size or shape of the objects [14].

Schiele et al. [8] classifies the visual people detection approaches into two major types: Sliding-window methods and part-based human body models. Sliding window methods scan the input images at each location and scale, classifying each window. Part based models, on the other hand, generate hyphothesis by evidence aggregation. The work by Schiele et al. [8] experimentally shows that part-based models outperform sliding window based models in the presence of occlusion and articulations however part-based approaches have a drawback which is the need of a high-resolution person existence in the images [8, 10]. [10] as a part-based model solves the object detection problem by developing a new multiscale deformable model. This detection system is modeled by parts of the object which are all trained in a discriminative training procedure. This object detection system achieves the best results in ten out of the twenty categories in PASCAL 2007 [10].

The literature includes a huge number of papers about visual tracking systems. An efficient feature-based tracking algorithm which searches and matches the candidate scale-invariant interest points in local neighbourhoods inside the 3D image pyramids is presented in [16]. A tracking algorithm using Scale Invariant Feature Transform (SIFT) based mean shift algorithm for object tracking by Zhou et al in [17]. Mean shift algorithm is used to conduct similarity search via color histograms for SIFT features over the region of interests. Liu et al. use not only color information but also motion information for modelling the environment. Liu et al. [18] integrates multiple cues into color-based Mean-Shift algorithm which is a fast algorithm for color blobs.

Tracking problem can be seen as a state estimation problem of dynamic systems. Ricon et al. [19] concentrates on the human legs benefiting from biomechanics constraint . The tracking is achieved by using a set of particle filters, and the bounding box of a person is tracked using Kalman Filter algorithm. A six-stick skeleton model of a pedestrian as a state space model is used in [20] and tracking is achieved by using particle filters. Tung et al. [21] uses an adaptive color based particle filter coupled with optical flow estimations for tracking. Lucas-Kanade Optical Flow algorithm is used by [22] for tracking human joints (head, torso and joints) and the joint movements are modeled with an articulate human stick model. Needham et al. [23] developed a multiple object tracking system using CONDENSATION algorithm. Each player

being tracked is independently fitted to a model, and the sampling probability for the group of samples is calculated as a function of the fitness score of each player.

[9, 24, 29] use tracking-by-detection technique that first detects the object of interests and tracks the targets. In this perspective, the performance of tracker is highly dependent to accuracy of the object detectors. Schiele et al. [8] examines the visual people detection approaches and compares their accuracies by re-implementing them.

The state-of-the-art techniques in the literature with their advantages and disadvantages are examined in this thesis. The background methods consist of four main parts:

- Motion detection
- Human detection
- Object tracking and
- Data association techniques for multi-target tracking.

2.1 Motion Detection

Motion is a valuable information for video processing. Hence, motion segmentation is generally used as a powerful preprocessing step to detect regions that can have potential targets in order to provide a focus of attention for later processing such as object recognition, tracking or action/activity understanding. The existing motion detection methods can be divided into three main categories; which are Temporal Differencing, Background Substraction and Optical Flow [5].

2.1.1 Temporal Differencing

Temporal differencing is a temporal information based approach for motion detection. This method classifies each pixels of the current frame as either background or foreground by first differencing the intensity values of each pixels of consecutive frames and then thresholding the resulting values. The difference values greater than a chosen threshold value are classified as foreground object where as the difference values smaller than the threshold values are

classified as background. This method is very simple and fast, however, it is very sensitive to the threshold value.

2.1.2 Background Substraction

In background substraction algorithms, the foreground objects are extracted using a reference image, called the *background model*. The simplest one of background algorithms is detecting moving regions in an image by taking the difference between the current image and the reference static background model. However it is very sensitive to illumination variations.

Recent backgorund substraction algorithms [5] focused on adaptive background models which eliminate the effects of illimunation variations and repetitive motion from clutter such as motion of leaves in a tree. Stauffer and Grimson [26] model each background pixel by multiple gaussian distribution, called Mixture of Gaussian (MoG). In this background model, if the current pixel value is matched with the Gaussian model, then the current pixel is decided as the background and the background model is updated with the current pixel value. An example of foreground-background segmentation using MoG is depicted in Figure 2.1.

(a) Original Image (b) Foreground Objects

Figure 2.1: Foreground extraction with Mixture of Gaussian Model

In this thesis, moving objects are detected using the Mixture of Gaussian model. After foreground-background segmentaion, some morphological operations and connected component analysis are applied to image in order to remove noises and for extracting foreground objects to be easily differentiated. This preprocessing phase is explained in detail in Chapter 3.

2.1.3 Optical Flow

Optical Flow uses characteristics of flow vectors of moving objects over time to detect moving regions in an image sequence [27]. One of the drawbacks of the optical flow method is that it extracts coherent motion under the assumption of brightness constancy and spatial smoothness.

2.2 Human Detection

Human detection is the task of distinguishing people from the background. An impressive progress has been accomplished in visual people detection techniques over the last few years although this problem has a challenging nature due to articulation, illumination variations, viewpoint variations etc. [8]. Schiele et al. The state-of-the art techniques for the sliding window approach for people detection are compared by Schiele et al. [8] and by Dalal et al. [11] and they are briefly described in the following subsections:

2.2.1 Haar Wavelets

Papageorgiou and Poggion [12] first proposed a trainable object detection system by using Haar Wavelets. The object class is represented in terms of local, oriented, multscale differences between adjacent regions which is computable by a Haar Wavelet transform which transforms the images from pixel space to wavelet coefficients. A large set of positive and negative object examples are trained by support vector machines to model the object class. The object classes for detections are faces, people and cars. For people class, three different types of wavelets (vertical, horizontal and diagonal) are used at the scale of 16 and 32 pixel. The visual boundary information is identified by the wavelets which are depicted in Figure 2.2. As depicted in this figure, the sides of the humans are represented by vertical wavelets, the head and the shoulders are represented by horizontal wavelets and the heads, feet, head and shoulders are represented by diagonal wavelets.

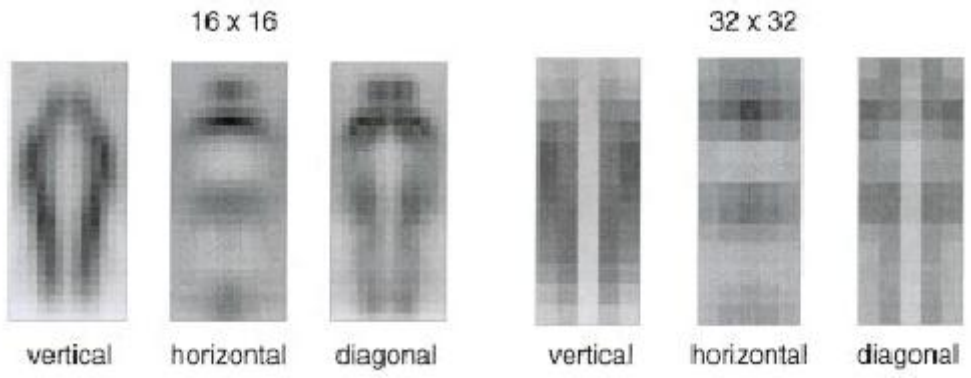


Figure 2.2: Visual boundary information of people identified by wavelets [12]

2.2.2 Histogram of Oriented Gradients

Histogram of gradients (HOG), proposed by Dalal and Triggs [11], has the idea that in the absence of precise knowledge of gradient or edge position, the local object appearance can be characterized better by the distribution of local intensity gradients or edge directions. HOG object detection method has a feature extraction chain resulting with a person / non-person classification. First the input image is normalized with a gamma correction, then the gradients are computed with an edge detector, then edge orientation histograms are formed (8x8 pixels) based on the orientation of the gradients. This orientation histograms are block normalized and the final vector which is formed by all normalized block histograms is used for person/non-person classification by using support vector machines (SVM). The HOG chain is depicted in Figure 2.3. Schiele et al. [8] and Dalal et al. [11] experimentally show that HOG method outperforms existing feature descriptors such as SIFT and Shape Contexts.

2.2.3 Shape Context

Shape Context has been originally proposed [28] for shape matching and object recognition. This approach uses a set of points sampled from the contour on the object and each of these points is associated with shape context descriptor which describes the arrangement of the rest of the shape with respect to that point. The shape context desriptor is invariant to shape deformations.

Leibe et al. [29] use shape context descriptor in Implicit Shape Model (ISM) framework for

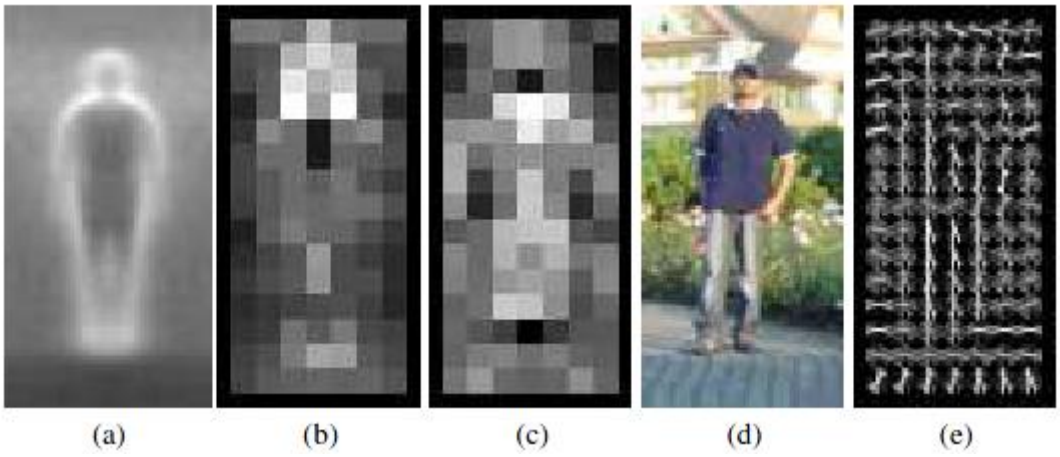


Figure 2.3: Silhoutte contours (the head, ahoulders and feet) are the cues [11]. (a) The average gradient image (b) Maximum positive SVM weight for each ”pixel” (c) Likewise for negative SVM weights (d) A test image (e) its computed R-HOG descriptor

pedestrian detection. The shape of object is extracted with a Canny edge detector. For a point *P* on the shape, a histogram is computed for the relative coordinates of the remaining points. This histogram is called shape context of *P*.

[29], [11] and [12] propose descriptors for human and pedestrian detections. In recent years, however, there is a tendency to find human body parts such as human heads[24] or upper body of humans [30] instead of finding the whole human silhouttes because the probability of occulusions for heads or upper body parts is less than the whole body [24] in crowded scenes.

2.3 Tracking

Tracking is the process of associating the detected object(s) throughout the video frames. Activity understanding and image interpretation, which are the processes of high level information extraction, are dependent on the performance of the tracking process. Tracking of objects is difficult due to complex object motion, partial/full occlusion of objects, scene illumination variations and noise in the images [14]. Tracking methods are classified into three major types [5, 14]: Point Tracking, Kernel Tracking and Contour Tracking.

2.3.1 Point Tracking

In point tracking the object is represented with points (Figure 2.4 (a)). The point representation is robust to the changes of rotation, scale and afine transformation[5]. In this method an external mechanism is required for object detection. The point correspondance problem exists due to misdetections and this problem can be solved with deterministic methods or statistical methods. Deterministic methods solve the correspondance problem constraining proximity, maximum velocity, and rigidity of object to be tracked. On the other hand, statistical methods use state-space representation of object including position, velocity, size and acceleration. In these methods, the pose estimation is usually done through Kalman Filter or Particle Filter [5].

2.3.2 Kernel Tracking

Kernel tracking is based on the computation of object motion where the object is represented by a geometrical region such as rectangular, ellipsodial or circular region (Figure 2.4 (b)). This tracking method is divided into two subcategories naming template model and appearance model [14, 5]. Template models are based on matching the object representation using similarity metrics [5] such as sum of squared differences (SSD), normalized crossed correlation and Battacharya coefficient. Optical Flow method is a popular kernel tracking method. On the other hand, multi view appearance model based kernel tracking uses an offline trained learning machine to overcome the appearance of object from different views[14].

2.3.3 Silhoutte Tracking

Objects generally have complex shapes rather than simple geometric shapes so silhoutte trackers extract the silhoutte of objects to be tracked and find the object region in each frame by using the object model generated from the previous frames. Silhoutte tracking methods are classified into two subcategories : shape matching and contour tracking. Shape matching methods are similar to template matching methods where an object silhoutte is searched in the current frame by using a similarity measurement. On the other hand, Contour tracking methods represent the object by its contour (Figure 2.4 (c), (d)) and find the object by iteratively evolving the contour in consecutive frames [14].

Silhoutte tracking methods has the advantage of handling a large variety of object shapes compared to kernel tracking methods. Another advantage of silhoutte tracking is the capability for dealing object split and merge for action understanding applications [14]. On the other hand, the tracking methods such as silhoutte tracking and kernel tracking are severely challenged by real-world scenes [9]. Real world scenes such as airports, streets and train stations are complex because they involve multiple people, complicated occlusions and backgrounds.

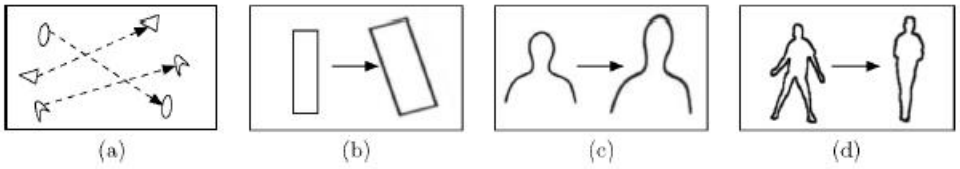


Figure 2.4: Different Tracking approaches [14] (a) Point Tracking (b) Kernel Tracking (c,d) Contour Tracking

2.4 Data Association Approaches for Multi-target Tracking

In multiple object tracking, the measurements need to be associated with the objects to be tracked as an additional task compared to single object tracking. This problem is called Data Association or Correspondence problem, depicted in Figure 2.5 , which needs to be solved before the measurements are applied to the filters like Kalman Filter or Particle Filter [14]. In order to tackle the correspondance problem [14], the simplest way is using the nearerst neighbour method, however it fails if the objects are close to each other. Multiple Hyphothesis Tracking (MHT) [31], Joint Probability Data Association Filter (JPDAF) [32] and Markov Chain Monte Carlo Data Association (MCMCDA) [15] techniques are widely used for data association problem which briefly are described in the following subsections:

2.4.1 Multiple Hyphothesis Tracking

The Multiple Hyphothesis Tracking (MHT) technique makes decision over multiple frames by deferring the correspondence decision [31]. The decision for forming a new track or removing an existing track is postponed until enough measurements are collected. Since multiple hypotheses are maintained at each frame for each object , the hypotheses grow exponentially over time and this results in an computational complexity. One of the advantages of the MHT

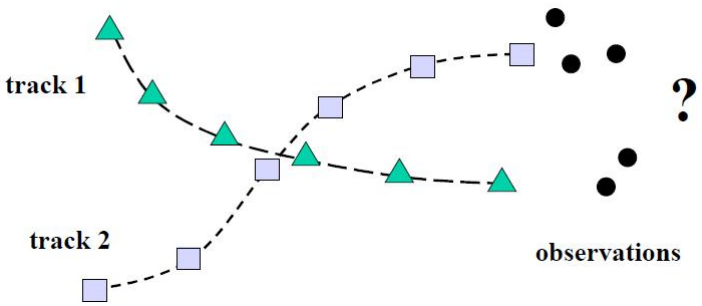


Figure 2.5: Which observations (measurements) belong to which track?

algorithm is the way it handles new tracks for objects entering the scene [14] which is suitable for surveillance applications.

2.4.2 Joint Probabilistic Data Association Filter

Joint Probabilistic Data Association Filter (JPDAF) algorithm is an extension of Probabilistic Data Association Filter algorithm which calculates the association probabilities for each validated measurement [32] at each time step by enumerating all possible associations instead of finding the best association between measurements and tracks. JPDAF uses only the measurements in the validation gate. Figure 2.6 shows two validation regions. The states of the objects are estimated by combining the measurement-to-track association probabilitiesB_{jk} where B_{jk} is the probability that j_{th} measurement extends the k_{th} track . JPDAF algorithm can handle the association of only fixed number of objects so it is not capable of handling new objects entering the scene[14]. On the other hand, JPDA has proved very effective in cluttered environments[15].

2.4.3 Markov Chain Monte Carlo Data Association

Markov Chain Monte Carlo Data Association (MCMCDA) algorithm, as the name suggests, uses Markov Chain Monte Carlo (MCMC) sampling instead of enumerating all possible associations. MCMC was first used to solve data association problem for a multi-camera traffic surveillance application containing hundreds of vehicles [33]. [15] introduces multi scan MCMCDA algorithm incorporating missing measurements, false alarms and handling varying tracks with real-time performance. [24] presents a multi-target tracking system which is

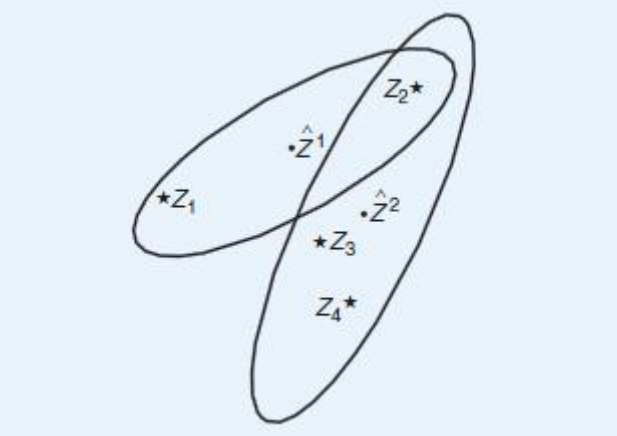


Figure 2.6: The two elliptical validation regions for two targets [32]

based on MCMCDA data association algorithm.

In conclusion, there are lots of techniques for both object detection and multi-object tracking in the literature. The smart surveillance applications need to recognize the object category in order to extract higher level information from the detections and tracking results which is called action/activity/ behaviour understanding in the literature [6]. In this thesis, we concentrate specifically on humans rather than objects as a general category and their motions.

The overall architecture is depicted in Figure 2.7. First, the background is modeled using Mixture of Gaussians method and region of interests are computed by foreground segmentation after applying shadow removal. This pre-processing step is used in many vision applications for better post-processing [14]. When a new frame is available, the humans are extracted by detecting their heads or upper bodies using self-trained detectors using HOG descriptors. The HOG descriptor is chosen for human detection since it gives the best results compared to other descriptors [8, 11]. The locations of the detections are represented by points due to its simplicty in crowded scenes [14] and they are associated with the tracks by using Markov Chain Monte Carlo Data Association algorithm (MCMCDA). Although there is no real-time requirement for this thesis, the MCMCDA is used for the data association problem for a possible need of real-time perpormance [15]. Finally, all detected humans are tracked with the Kalman Filter which is extensively used in vision community in order to track targets represented by points in state space [14].

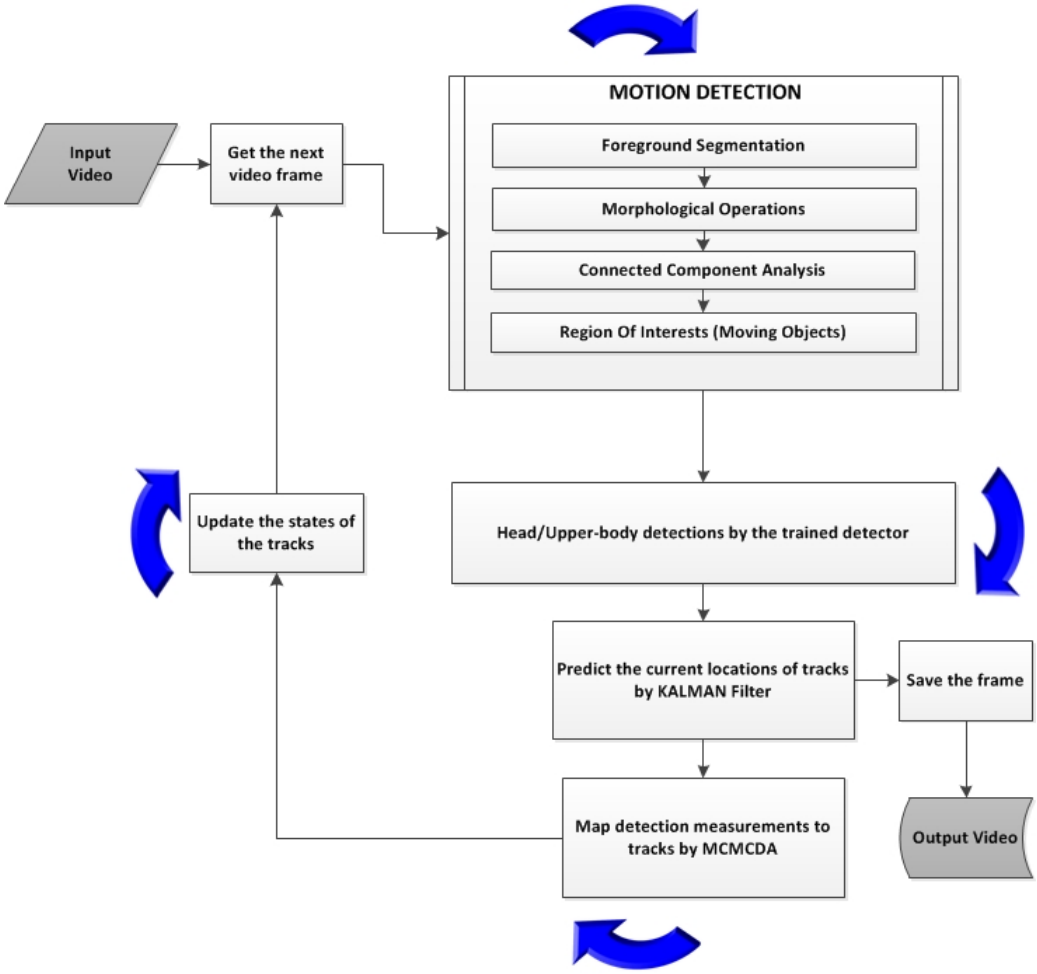


Figure 2.7: The overall architecture for multi-human tracking

The implemented system has an iterative process. When a new frame is fetched from the input video, a pre-processing step is applied. After the head/upper body detections, measurements are associated to each track and tracks are corrected. The main logic of the system is maintaining the state of each track. The number of the tracks is dynamically handled by MCMCDA. This process continues until all frames run out.

CHAPTER 3

MOTION DETECTION

The detection of moving objects in the scene is the first step of surveillance applications [14]. The performance of the pre-processing step directly affects the performance of the overall architecture since the human head detection and upper body detection steps are highly dependent on this step. The foreground is segmented by modelling the background pixels using Mixture of Gaussians (MoG) technique.

When a new video frame is available, the foreground objects are extracted by MoG. The noises are removed by applying some morphological operations. The region of interests (ROIs) for detecting human heads/upper bodies are extracted by applying the connected component analysis method in order to reduce the time for detections by focusing only these ROIs. The pre-processing step is depicted in Figure 3.1.

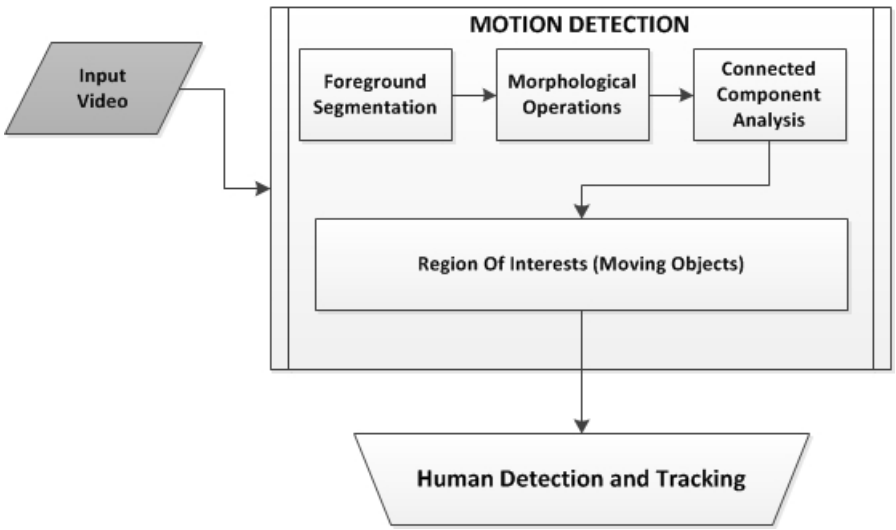


Figure 3.1: The pre-processing step of the overall architecture

3.1 Background Modeling

In this thesis, the foreground objects are extracted by using Mixture of Gaussians (MoG) method. The MoG proposed by Stauffer and Grimson [26] models every pixel in the image with a mixture of Gaussian distributions. The recent history of intensity value of each pixel $X_1, ..., X_t$ is modeled by a mixture of K Gaussians. The probability of observing the current pixel value is

$$P(X_t) = \sum_{i=1}^{K} \omega_{i,t} * \eta(X_t, \mu_{i,t}, \Sigma_{i,t}) \tag{3.1}$$

where the number of distributions are represented by K, weight associated to the i^{th} Gaussian at time t is represented by $\omega_{i,t}$, the mean and the covariance matrix of the i^{th} Gaussian at time t are represented by $\mu_{i,t}$ and $\Sigma_{i,t}$ respectively. η is the multivariate Gaussian distribution whose probability density function is given in equation (3.2).

$$\eta(X_t, \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(X_t-\mu)^T\Sigma^{-1}(X_t-\mu)} \tag{3.2}$$

The number of Gaussian distribution, K, is proposed to be 3 to 5 by Stauffer and Grimson [26]. [26] assumes independent variances for computational reasons. Therefore the covariance matrix is in the form:

$$\Sigma_{i,t} = \sigma_{i,t}^2 I \tag{3.3}$$

The parameters for gaussian distributions $\omega_{i,t}$, $\mu_{i,t}$ and $\Sigma_{i,t}$ are initialized by the K-means algorithm for real time requirements [26].

When a new frame is captured, a match test is made for each pixel. If a match is found with one of the K Gaussians, then the pixel is classified as a background pixel, otherwise the pixel is classified as a foreground pixel. The match test for a pixel is made by using the Mahalanobis distance

$$\sqrt{(X_{t+1} - \mu_{i,t})^T\Sigma^{-1}(X_{t+1} - \mu_{i,t})} < k\sigma_{i,t} \tag{3.4}$$

where k is a constant threshold equal to 2.5 [26]. After the match test, if a match is found with one of the K Gaussians, then the parameters $\omega_{i,t+1}$, $\mu_{i,t+1}$ and $\Sigma_{i,t+1}$ are updated for the matched components and only $\omega_{i,t+1}$ is updated for unmatched components. If a match is not found with one of the K Gaussians, the least probable distribution is replaced with a new one with its new parameters.

An example modelling is depicted in Figure 3.2. This figure shows the shadows with gray color.

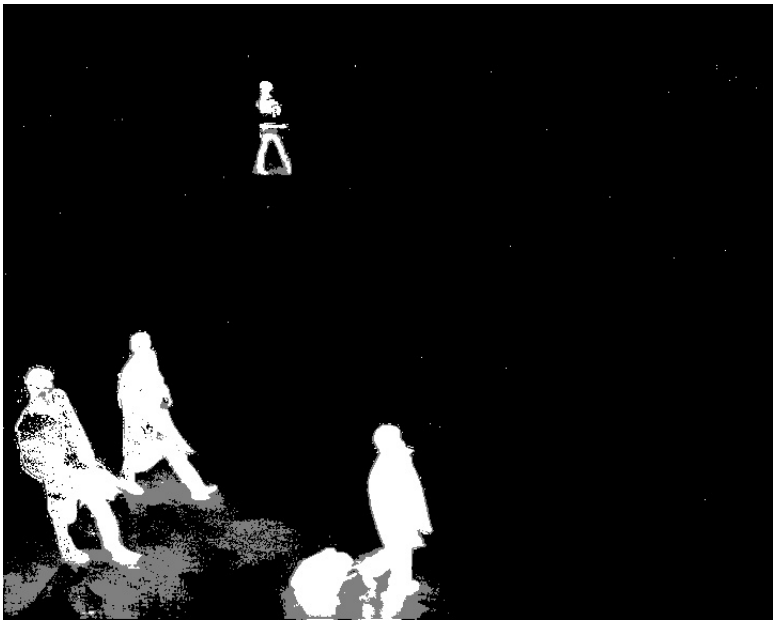


Figure 3.2: Foreground-Background image with shadow

The shadows are removed for more precise region of interests. Smaller region of interests provides less false alarms for detection. The scene with shadow removal is depicted in Figure 3.3.

3.2 Morphological Operations

The next pre-processing step is the application of Morphological operations. *Morphology* is a set of image processing operations that process images based on shapes. The set of operations are generally used for removing noises in a binary image before a connected component labeling algorithm is applied. The inputs for morphological operations are the input image

Figure 3.3: Foreground-Background image without shadow

and the structuring element (also called kernel) and the output is an output image of the same size. The structuring elements represents a shape such as rectangular or circular shapes.

The structuring element is shifted through the input image in such a manner that the center pixel of the mask is matched with each pixel location of the input image. This is depicted in Figure 3.4 with a sharpening structuring element.

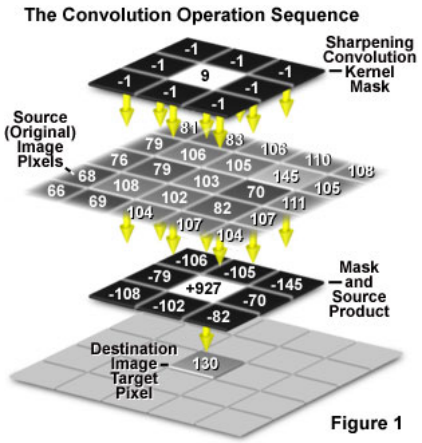


Figure 3.4: Convolution of the input image [34]

The basic morphological operations are the erosion and dilation operations which are generally used for removing noise and isolating individual elements. The combination of erosion

and dilation operations are called openning and closing operations. The erosion and dilation operations are explaind in the following subsections.

3.2.1 Erosion

Erosion completely removes objects smaller than the structuring element and shrinks greater than structuring elements for binary images. While structuring element is shifted over the binary image, if the structuring element is the same as region in the input image, then the pixel in the output image becomes binary 1-pixel, otherwise becomes binary 0-pixel. The structuring element for erosion operation used in this thesis is given in (3.5).

$$SE_{Erosion} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 0 \end{pmatrix} \tag{3.5}$$

3.2.2 Dilation

Dilation adds pixels to the boundaries of objects in an image. While structuring element is shifted over the binary image, if the center of structuring element touches a binary 1-pixel, the entire structuring element is logically "'OR"'ed with output image whose all pixels are initialized to a binary 0-pixel. The structuring element for dilation operation used in this thesis is given in (3.6).

$$SE_{Dilation} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \tag{3.6}$$

3.3 Connected Component Analysis

Connected component analysis is the labeling operation in which the value of each foreground pixel is labeled with its component label and used for detecting the connected regions. The connected components of Figure 3.2 are depicted in Figure 3.5 and the connected components

of Figure 3.3 are depicted in Figure 3.6.

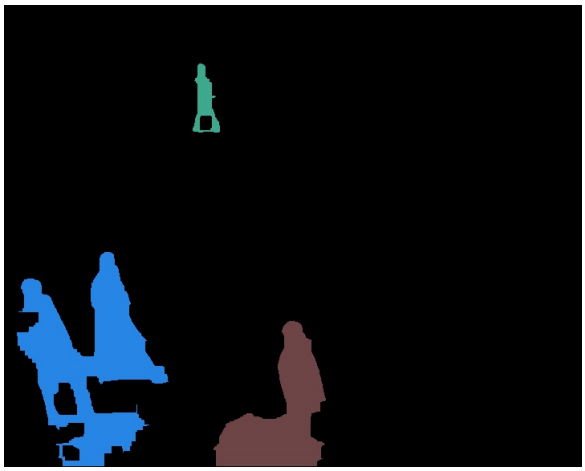


Figure 3.5: The connected components with shadow

Figure 3.6: The connected components without shadow

After connected componets are labeled, region of interests (ROIs) are generated using the boundaries of components by additionaly adding margins for two axes. These ROIs are used for human body part detection with a detector trained in Chapter 4. They are depicted in Figure 3.7 without shadow removal and Figure 3.8 with shadow removal.

Figure 3.7: The ROIs generated without applying shadow removal

Figure 3.8: The ROIs generated with applying shadow removal

Figure 3.8 involves more precise region of interests compared to Figure 3.7. More precison provides less false alarms and less computational time. However false shadow detections can cause losing regions of interests for further processing. While Figure 3.7 has 3 ROIs, Figure 3.8 has 2 ROIs. The effects of the shadow removal will be discussed in Chapter 4.

CHAPTER 4

HUMAN BODY PART DETECTION

Real world scenes such as airports, streets and train stations are complex because they involve many people, complicated occlusions and cluttered backgrounds. Although complex real world scenes exist, human detectors are able to locate pedestrians with recent developments in computer vision. In Chapter 2, the human detectors were briefly described. An object detection mechanism is required for every tracking method. The performance of the object detector will directly affect the performance of object tracker.

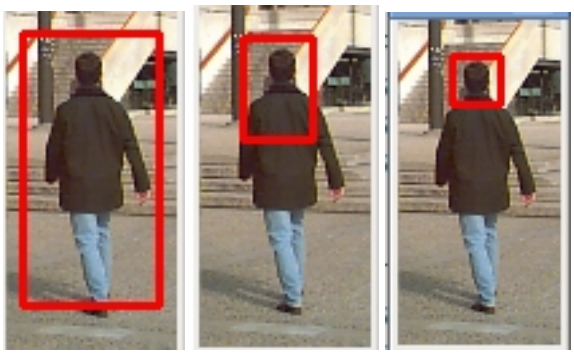
[8] and [9] claim that combining detection and tracking is a promising direction for crowded real world scenes. [24] uses human head detections for tracking and [30] uses human upper body part detections for tracking and action understanding.

In this chapter, first the HOG descriptor proposed by Dalal and Triggs [11] is explained. Then, the experimental work for human head detections and upper body detections done throughout this study will be explained. Finally the performance of the trained detector will be discussed by evaluating the effects of the parameters in the training process.

4.1 HOG Descriptor for Human Detection

Human detection is a challenging task due to a wide variety of articulated poses of humans, variable appearance, complex backgrounds, occlusions, illumination variance and different scales of humans in images. Dalal et al. [11] show experimentally that locally normalized Histogram of Oriented Gradients (HOG) descriptor outperforms existing feature sets including wavelets, SIFT Descriptors and shape contexts. The basic idea of HOG is providing the distribution of local intensity gradients and edge directions to characterize the object appear-

ance.



(a) Human Body Detection (b) Human Upper Body Detection (c) Human Head Detection

Figure 4.1: Human Part Detections

The HOG based human detection in images has two phases: Learning phase (Figure 4.2 -(a)) and Detection phase (Figure 4.2 -(b)). In the learning phase a binary classifier is trained to provide the object/non-object decision for image windows. In the detection phase the pre-trained binary classifier is used by multi-scale scanning of every location of the test image [9].

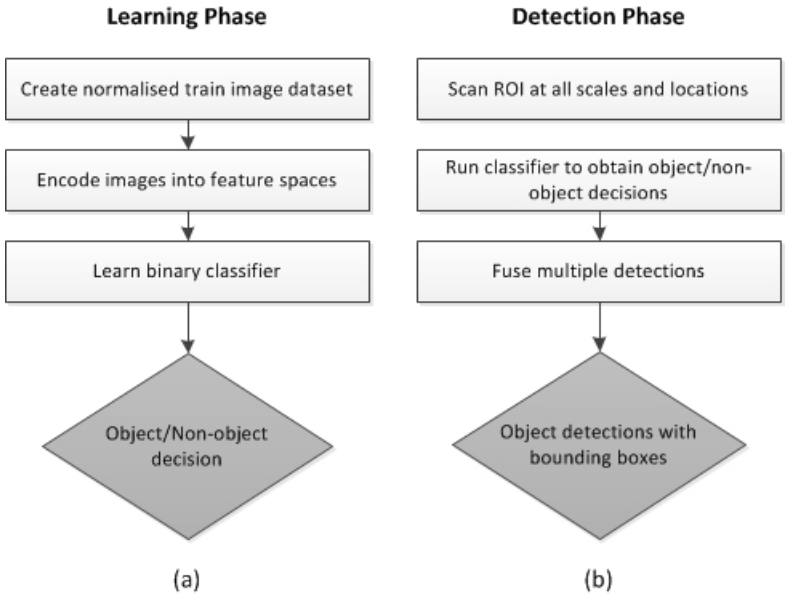


Figure 4.2: Overall Human Detection Architecture a) Learning Phase b) Detection Phase [11]

Dalal et al. [11] first created a new dataset containing positive examples with ideally one human in it and negative examples with no human. The binary classifier is trained for

person/non-person decision by using this dataset. Dalal et al. [11] used Linear SVM as a binary classifier for the training phase due to its reliable convergence, handling large datasets and robustness for varying feature sets and their parameters. The feature vectors extracted from the image window are fed into SVM classifier for learning.

The experimental work done by Dalal et al. [11] is about the feature extraction process (Figure 4.3) which transforms the pixel space to feature space. The proposed descriptor [11] is based on the distribution of local intensity gradients and edge orientations which characterizes the object appearance and shape.

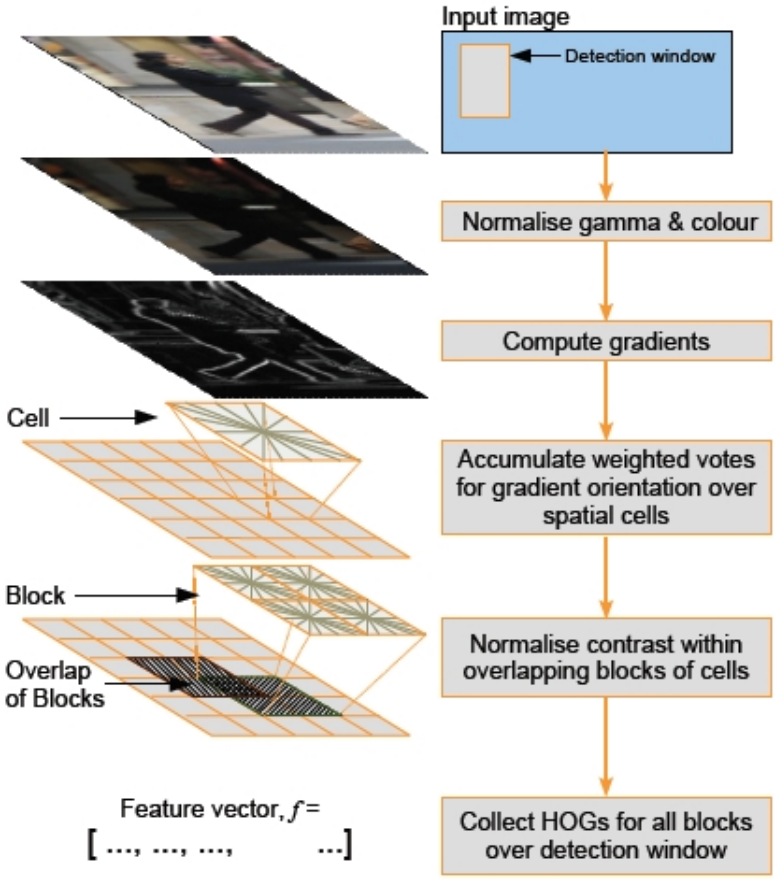


Figure 4.3: HOG feature extraction chain [11]

The feature extraction steps which are used in both learning and detection phases(depicted in Figure 4.3)) are briefly described below:

- **Global Normalization:** Gamma correction (normalization) is applied for each color

channel either using square root function or log function. This normalization provides to reduce the effects of local shadowing and illumination variations. The formula is given below:

$$V_{out} = AV_{in}^{\gamma} \tag{4.1}$$

where A is a constant scalar and the input and output values of matrices are non-negative real values.

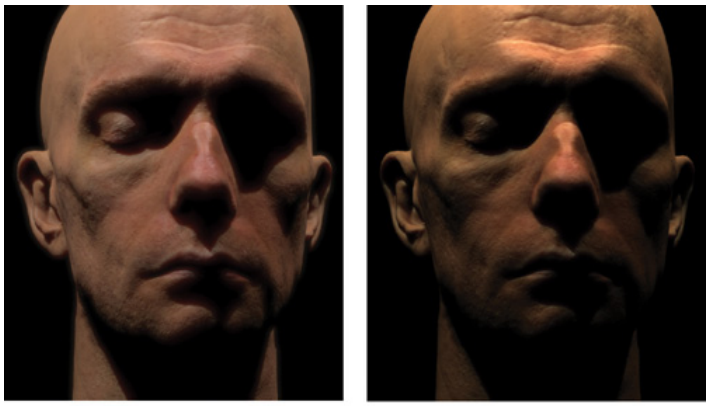


Figure 4.4: Application of gamma normalization. (a) Original image (b) Normalized image with $\gamma < 1$

- **Gradient Computation:** The edges are computed for each channel and the locally dominant color channel is used. This provides color invariance.
- **Forming Orientation Histograms:** The image window is divided into regions called ”cells” and for each cell a local 1D histogram of gradient is formed. This histogram consists of orientation bins. The gradient magnitudes vote one of the orientation bin of the orientation histogram.
- **Local Normalization:** A group of cells are combined and called as “blocks”. The blocks are formed by a measure of local histogram “energy”. Since the blocks are overlapped, the cells appear several times in the final feature vector.
- **Forming feature vector:** The feature vector is formed from all blocks covering the detection window.

In the detection phase the test image is scanned at each scale and location. At each detection window the pre-trained binary classifier is run to produce object/non-object decision. Dalal et

al. [11] discuss the effects of some parameters such as the number of orientation bins, gamma correction, gradient scale, normalization methods and window size in their study.

4.2 Human Head and Upper Body Training

In this thesis, OpenCV and SVMLight are used for HOG descriptor computation and binary classification (person/ non-person decision) is done on a Linux machine with 2.10 GHz CPU. The datasets used for both human head detections and human upper body detections are normalized such that heads and upper bodies are centered at the image.

During the training process, the effects of some parameters are examined and the detection rate is recorded. The number of positive images are doubled with their reflections for both human heads and human upper bodies. The parameters are listed below:

- Detection window size
- Number of orientation bins (spaced over 0 - 180)
- Effect of overlap
- Gamma Correction
- Re-training process (Hard training)

The SVM parameters for both human head and human upper body training processes are same as [11] with linear kernel and trade-off between training error and margin.

4.2.1 Human Head Training

For human head training, the dataset which belongs to "Oxford University - Active Vision Group" is used for positive training set. This dataset has 1836 images that are all 24x24 pixels with the center 16x16 pixels containing the head (Some of them are depicted in Figure 4.5). As a negative training dataset 100 human-free images are used.

The only window size for human head training process is 24x24. So there is no window size variation for this experimental training process. The results obtained by changing the

Figure 4.5: Human head images for training

parameters such as the number of orientation bins, size of overlap, gamma correction are explained below:

The trained HOG Descriptor is tested with both the MIT Pedestrian dataset and the INRIA Person dataset including only positive data. On the other hand, the trained HOG Descriptors are tested separately with the INRIA Negative person dataset in order to observe the false positives. For each image in the 600 images of INRIA Negative dataset, 10 random image patches are cropped which makes 6000 image patches in total.

The Number of orientation bins:

In this training procedure the effects of the number of orientation bins are observed. The cell size used in this training is 4x4 pixels for a given training image having a 24x24 pixels size and 2x2 cells size is used for descriptor blocks. 75% block overlapping is used. The experimentally observed numbers of orientation bins are 6, 9, 12, 18.

The results for positive dataset are given in Table 4.1 for both test datasets.

Table 4.1: Effect of Number of Orientation Bins for Human Head Training - Recall

#	Recall(MIT)	Recall(INRIA)
6	83%	79%
9	87%	82%
12	82%	78%
18	77%	76%

The rate of false positives are given in Table 4.2.

Table 4.2: Effect of Number of Orientation Bins for Human Head Training - False Positives

#	False Positives(INRIA)
6	505 / 6000
9	523 / 6000
12	516 / 6000
18	522 / 6000

Effect of block overlapping rate:

In this training procedure the effects of overlapping of descriptor blocks are observed. The cell size used in this training is 4x4 pixels for a given training image having a 24x24 pixels size and 2x2 cells size is used for descriptor blocks. The number of bins for orientation histograms is 9. The experimentally observed overlapping rates are 0% , 25%, 50%, 75%.

The results for the positive dataset are given in Table 4.3 for both test datasets.

Table 4.3: Effect of Block Overlapping Rate for Human Head Training - Recall

%	Recall(MIT)	Recall(INRIA)
0	76%	71%
25	82%	77%
50	84%	79%
75	87%	82%

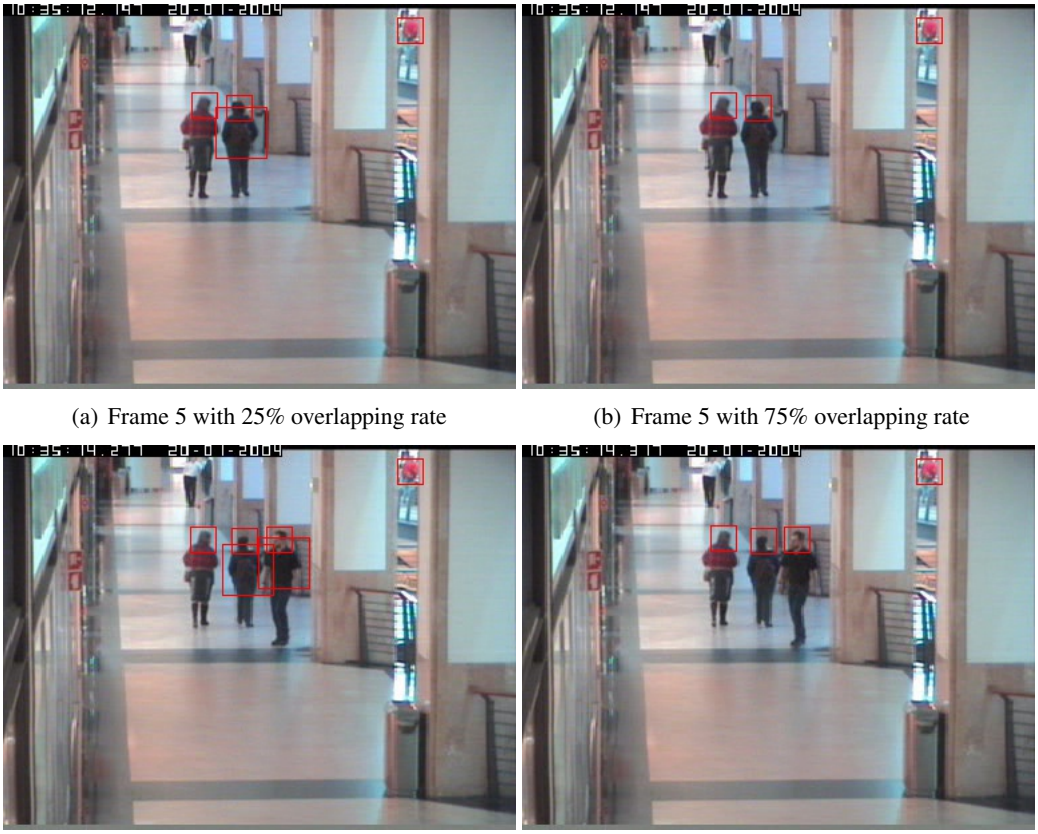
The rate of false positives are in Table 4.4 below.

Table 4.4: Effect of Block Overlapping Rate for Human Head Training - False Positives

%	False Positives(INRIA)
0	584 / 6000
25	551 / 6000
50	547 / 6000
75	523 / 6000

The results of the increasing block overlapping rate are depicted in Figrue 4.6. In Figure 4.6 b and 4.6 d, the false positive detections are ommited with 75% block overlapping rate.

Gamma Correction: In this training procedure the effects of applying gamma correction are observed. The cell size used in this training is 4x4 pixels for a given training image having a 24x24 pixels size and 2x2 cells size is used for descriptor blocks. 75 % block overlapping is



(a) Frame 5 with 25% overlapping rate

(b) Frame 5 with 75% overlapping rate

(c) Frame 58 with 25% overlapping rate

(d) Frame 58 with 75% overlapping rate

Figure 4.6: Effect Of Block Overlapping on Human Detection

used. The experimentally applied gamma functions are square root and log functions. They have very ignorable effects, they change the recall rate very modestly.

Hard-training: In [11] negative training dataset is re-trained in order to reduce the false positive rates. This technique is applied here with initial 100 images by exhaustively searching them. However this method increases the false negative rate.

4.2.2 Human Upper Body Training

For human upper body training, 32x32 and 32x48 image patches of MIT Pedestrian dataset consisting 924 images of size 128x64 pixels are used for the positive training set. As a negative training dataset 100 human-free images are used. The window sizes of 32x32 pixels (Figure 4.7) and 32x48 pixels (Figure 4.8) are trained separately. The results obtained by changing the parameters such as the number of orientation bins, rate of block overlapping,

gamma correction are explained below.

The trained HOG Descriptor is tested with INRIA Person dataset both including positive dataset and negative dataset. For each image of 600 images of INRIA Negative dataset, 10 random image patches are cropped which totally equals to 6000 image patches.

Figure 4.7: 32x32 px upper body images

Figure 4.8: 32x48 px upper body images

Number of orientation bins:

In this training procedure the effects of the number of orientation bins are observed. The cell size used in this training is 8x8 pixels for a given training image having size of 32x32 and 32x48 pixels. 2x2 cells size is used for descriptor blocks. 75 % block overlapping is used. The experimentally observed numbers of orientation bins are 6, 9, 12, 18.

The results for positive dataset are given in Table 4.5 for both 32x32 and 32x48.

The rate of false positives are given in Table 4.6.

Effect of block overlapping rate: In this training procedure the effects of overlapping of

Table 4.5: Effect of Number of Orientation Bins for Human Upper Body Training - Recall

#	Recall(32x32)	Recall(32x48)
6	77%	80%
9	79%	82%
12	78%	81%
18	75%	78%

Table 4.6: Effect of Number of Orientation Bins for Human Upper Body Training - False Positives

#	F.P.(32x32)	F.P.(32x48)
6	323 / 6000	236 / 6000
9	301 / 6000	239 / 6000
12	312 / 6000	245 / 6000
18	336 / 6000	235 / 6000

descriptor blocks are observed. The cell size used in this training is 8x8 pixels for a given training image having size of 32x32 and 32x48 pixels and 2x2 cells size is used for descriptor blocks. The experimentally observed overlapping rates are 0%, 25%, 50%, 75%.

The results for positive dataset are given in Table 4.7 for both 24x32 and 32x48 pixel images.

Table 4.7: Effect of Block Overlapping Rate for Human Upper Body Training - Recall

%	Recall(32x32)	Recall(32x48)
0	73%	76%
25	74%	78%
50	77%	81%
75	79%	82%

The rates of false positives are given in Table 4.8.

Gamma Correction: In this training process the application of gamma correction has the same effect that head training process has.

Hard-training: In this training process again hard-training has the same effects that head training has. While the false positive rate decreases, the false negative rate increases.

Table 4.8: Effect of Block Overlapping Rate for Human Upper Body Training - False Positives

%	F.P.(32x32)	F.P.(32x48)
0	361 / 6000	278 / 6000
25	339 / 6000	267 / 6000
50	327 / 6000	248 / 6000
75	301 / 6000	239 / 6000

4.3 Discussion

For human detection both the head detection and the upper-body detection approaches have their advantages and disadvantages. The head detection with trained HOG Descriptor is faster than upper-body detection due to the window size and the HOG Descriptor size. However the head detection comes with higher rates of false positives compared to the upper body detection. An example that explains this situation is depicted in Figure 4.9 and Figure 4.10;

Figure 4.9: Head Detections

This experimental study has similar outcomes compared to Dalal's et al. study [11]. Dalal et al. claim that fine orientation coding and higher rate of overlapping of spatial blocks are very critical for good performance for the whole body detection. In this study we have similar outcomes for both head detection and upper-body detection. When the rate of spatial block overlapping is increased, the size of the HOG descriptor increases and the detection time

Figure 4.10: Upper Body Detections

decreases.

Another important point is the occlusion problem. Our experimental work shows that detecting human heads or human upper bodies is more robust than detecting human bodies from occluded humans since the feature extracted from the image is not enough to discriminate the human bodies from occuluded humans. Figure 4.11 shows this situation.

(a) Upper Body Detections (b) Full Body

Figure 4.11: Comparison of Upper Body Detections and Full Body Detection

Hard-training technique is used in order to decrease the false-positive rate which is the main problem in human detection [9]. This false-positive detections are the noisy measurements for

the tracking problem and they will be discarded at the tracking step by the application of the filtering method i.e. Kalman Filter. However if a human head or upper-body cannot be found by the detector although it exists, this is a problem for tracking because it cannot be solved in any step. So in order to form a final HOG descriptor our strategy has been decreasing the false-negative rates.

The head and upper body detections from MIT Pedestrian dataset using trained HOG Detector are depicted in Figure 4.12 and Figure 4.13 respectively.

Figure 4.12: Images containing detected human heads

Figure 4.13: Images containing detected human upper body parts

The upper body detections by the trained HOG Detector for images containing humans at different scales are depicted in Figure 4.14, Figure 4.15, Figure 4.16 and Figure 4.17.

Figure 4.14: Upper Body Detections with different scales from INRIA Dataset-1

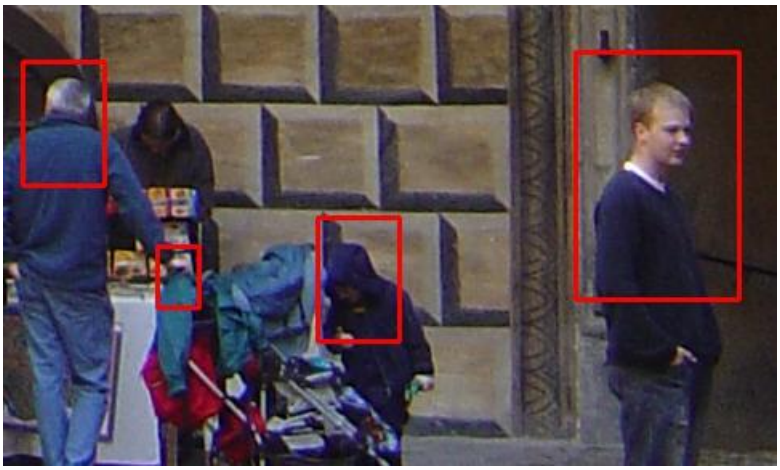


Figure 4.15: Upper Body Detections with different scales from INRIA Dataset-2

Figure 4.16: Upper Body Detections with different scales from INRIA Dataset-3

Figure 4.17: Upper Body Detections with different scales from INRIA Dataset-4

The upper body detections trained by the HOG Detector in a video containing occluded walking humans are depicted in Figure 4.18. The human detector proposed by Dalal and Triggs [11] was not able to discriminate the occluded humans in this scenario. Figure 4.11 shows both upper body detections and full body detection result in the same scene.

(a) Frame 61 (b) Frame 68 (c) Frame 74

(d) Frame 85 (e) Frame 86 (f) Frame 87

(g) Frame 88 (h) Frame 101 (i) Frame 128

Figure 4.18: Upper body detections for video with occlusions

The final human head detector that generates measurements for the tracker is formed with 24x24 window size, 9 orientation bins, 75% block overlapping rate. On the other hand, the final human upper body detector is formed with 32x48 window size, 9 orientation bins and 75% block overlapping rate. The training processes of final detectors do not involve the hard training phase due to the high rate of false negatives. These detectors' computational performance is measured with a 720x576 resolution video from PETS2006 Dataset-S4 [39]. This video has 3020 frames. Table 4.10 gives the detection time percentage for a single frame.

These rates are calculated by the division of total processing time for the video over the 3020 frames. Application of shadow removal method gives better processing time due to more precise region of interests. Small detection window size constitutes small HOG Descriptor size for the detection. Small HOG descriptor size takes less processing time for a frame.

Table 4.9: Time performance comparisons

sec/frame	Head	Upper Body	Full Body
Shadow Removal	0.537	0.881	1.84
No Shadow Removal	0.769	1.067	1.912

The precision and recall values of the trained head detector and upper body detector are compared with the full body detector with INRIA person test dataset. This dataset contains 741 images. 288 images of this dataset involves human in it.

Table 4.10: Precision and Recall rates for the detectors

%	Head	Upper Body	Full Body
Precison	72.8%	79.7%	85.5%
Recall	77.9%	82.3%	88.7%

Precision value in head detection is smaller compared to the other detectors because of the high rates of false positives. The recall values show more succesfull performance due to the less rates of false negatives. According to both metrics, full body detector provides more accurate detections but the trained upper body detector is almost same as the full body detector in terms of the accuracy. The circular shape of the human head leads to more false positive rates.

CHAPTER 5

MULTI HUMAN TRACKING

Multiple-Target Tracking problem is distinguished from Single-Target Tracking problem with data association problem which is the problem of mapping the measurements to the tracks especially in a cluttered environment. In this thesis, Multi-Scan version of Markov Chain Monte Carlo Data Association algorithm is used rather than Joint Probabilistic Data Association Filter (JPDAF) and Multiple Hyphothesis Tracking (MHT) algorithms described in Chapter 2. Oh et al. [15] proves that single-scan MCMCDA approximates JPDAF when the number of the targets to be tracked is fixed. On the other hand, Oh et al. [15] presents the multi-scan version of MCMCDA for unknown numbers of targets proving that Multi-Scan MCMCDA has a better performance compared to MHT for a large number of target in a dense environment and high false alarm rates.

5.1 Multiple-Target Tracking Problem Formulation

Let K number of objects appear in surveillance region R with duration T. Each object appears in R at t_i^k and disappears at t_f^k where t_i^k and t_f^k are unknown. Let $F^k : R^{n_x} \rightarrow R^{n_x}$ be the discrete-time dynamics of object k where n_x is the dimension of the state-space, and the current state of the object is denoted with x_t^k at time t. The next state of the object k is computed by:

$$x_{t+1}^k = F^k(x_t^k) + w_t^k, \qquad \text{for } t = t_i^k, ..., t_f^k - 1 \tag{5.1}$$

where $w_t^k \in \mathbb{R}^{n_x}$ are white noise processes. The measurements are observed with the probability p_d meaning that the object is not detected with probability $1 - p_d$. Let y_t^j be the j-th

observation at time t for j = 1,..,n_t where n_t is the number of total observations at time t.

The multiple target tracking problem is to estimate the number of tracks (K), t_i^k, t_f^k for k = 1,...,K given the measurements. These measurements are generated from the HOG detections explained in Chapter 4. The multi-object tracker finds a set of tracks, ω, for K objects called partition (Figure 5.1) through tracking period T from the set of all measurements Y = Y_t, t = 1,...,T where $Y_t = y_t^j$: j = 1,...,n_t at time t.

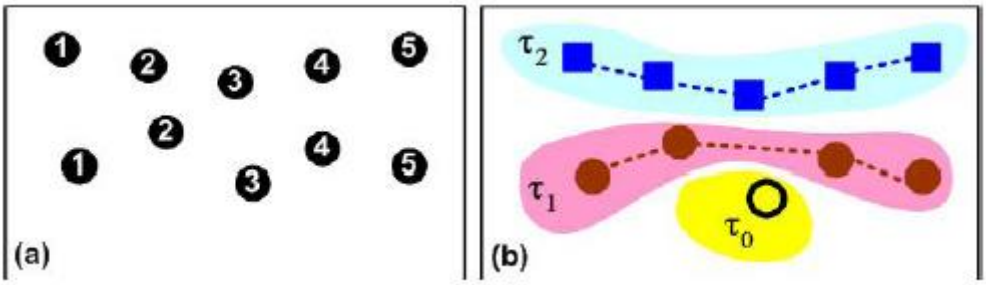


Figure 5.1: (a) The observed measurements at time t. (b) An example of a partition ω of Y where τ_0 represents false-alarm [15]

As a result, the posterior of ω can be formulated by an equation

$$P(\omega \mid Y) \; \alpha \; P(Y \mid \omega) \prod_{t=1}^{T} p_z^{z_t}(1 - p_z)^{c_t}(1 - p_d)^{g_t} \lambda_b^{a_t} \lambda_f^{f_t} \tag{5.2}$$

where

- $P(Y \mid \omega)$ is the likelihood of measurements Y given ω,
- λ_b is the birth rate of new objects per unit time,
- λ_f is the false alarm rate per unit time,
- a_t is the number of new targets
- c_t is the number of targets from t - 1
- d_t is the number of actual target detections at time t
- g_t is the number of undetected targets that can be found by $g_t = c_t + a_t - d_t$

5.2 MCMCDA for Multiple Target Tracking

Markov Chain Monte Carlo methods are algorithms for sampling from a probability distribution π, which is the density of interest, on a space Ω constructing a Markov Chain M with states where Ω is the collection of partitions Y (Figure 5.1 - b).

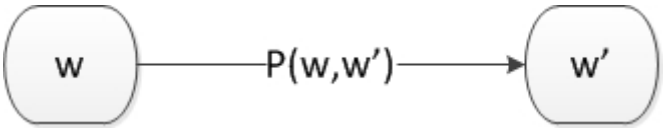


Figure 5.2: Transition probability from state ω to $\omega^{'}$

In this thesis, Metropolis-Hastings algorithm is used as an MCMC method which is a Random Walk algorithm [35]. This algorithm generates a random "move" using a proposal density and provides a mechanism for accepting/rejecting this generated "move".

In Metropolis-Hastings algorithm, a new state is proposed following the proposal distribution q(ω,$\omega^{'}$) to move from state ω. The move is accepted, in which case the chain moves, with an acceptance probability A(ω,$\omega^{'}$) where

$$A(\omega, \omega^{'}) = min\left(1\ ,\ \frac{\pi(\omega^{'})q(\omega^{'}, \omega)}{\pi(\omega)q(\omega, \omega^{'})}\right) \tag{5.3}$$

otherwise the current state is kept that is the move is rejected. In other words, whether or not the move is accepted or rejected depends on the acceptance probability. The $\pi(\omega)$ is defined by

$$\pi(\omega) = P(\omega \mid Y) \tag{5.4}$$

where $P(\omega \mid Y)$ defined in Equation 5.2

The transition probability is given by

$$P(\omega \mid \omega^{'}) = q(\omega, \omega^{'})A(\omega, \omega^{'}) \tag{5.5}$$

where ω is the current state and $\omega^{'}$ is the proposed state which is depicted in Figure 5.2.

5.3 Multi-Scan MCMCDA Algorithm

The Multi-Scan MCMCDA algorithm has a significant performance compared to MHT algorithm for unknown number of targets [15] . The inputs to the algorithm are the set of all measurements denoted with Y, the number of samples n_{mcmc} , the initial state w_{init} and the X which is a bounded function $X : \Omega \rightarrow \mathbb{R}^n$. The X and $\hat{X}$ are used as a metric which estimates the minumum mean square error (MMSE).

Algorithm 1 Multi-Scan MCMCDA algorithm

Input: Y, n_{mc}, $X : \Omega \rightarrow \mathbb{R}^n$

Output: $\hat{\omega}$, $\hat{X}$

$\omega = \omega_{init}$; $\hat{\omega} = \omega_{init}$; $\hat{X} = 0$;

for $n = 1 \rightarrow n_{mc}$ **do**

propose ω' based on ω

sample U from Unif[0,1]

$\omega = \omega'$ **if** $U < A(\omega,\omega')$

$\hat{\omega} = \omega$ **if** $P(\omega \mid Y) \,/\, P(\hat{\omega} \mid Y) > 1$

$\hat{X} = \frac{n}{n+1}\,\hat{X} + \frac{1}{n+1}\,X(\omega)$

end for

Multi-Scan MCMCDA algorithm depicted in Algorithm 1 proposes a new partition $\omega^{'}$ by choosing a random move from the proposal distribution. The proposal distribution is based on some assumptions. When there is no track, the only proposed move is the birth move. If there exist only one track, a merge or track switch move will not be proposed.

Multi-Scan MCMCDA has a proposal distribution consisting 5 fundamental moves (totally 8 moves) which are depicted in Figure 5.3.

- Birth/Death move pair
- Split/Merge move pair
- Extension/Reduction move pair
- Track update move
- Track switch move

The number of tracks denoted by K is updated depending on the type of the proposed move. For a birth move and split move, the number of tracks denoted by K is incremented by one and for a death move and merge move, K is decremented by one.

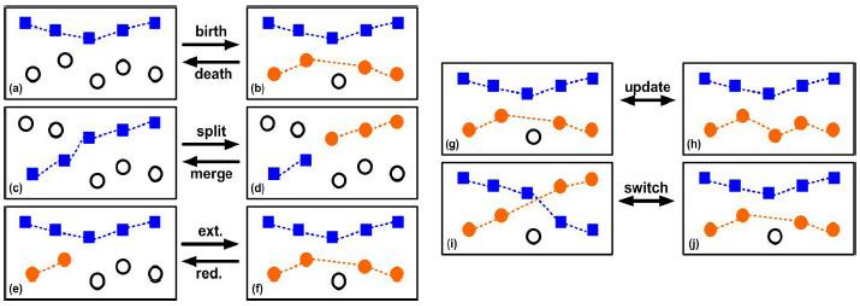


Figure 5.3: Different types of moves [15]

The matlab simulation results of Multi-Scan MCMCDA algorithm for Figure 4.18 in Chapter 4 are depicted in Figure 5.4. Since head detections with the trained HOG detector generates many false positives, the Multi-Scan MCMCDA algorithm is used with human upper body detections.

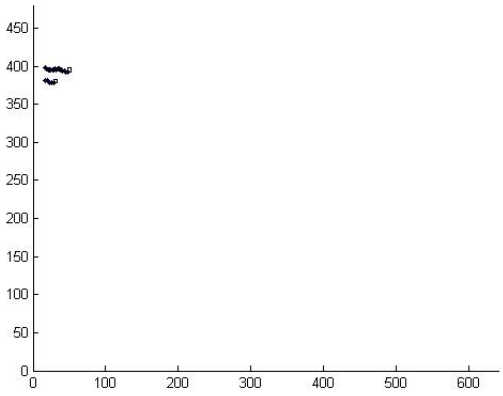
5.4 State Estimation

After measurements are assigned to current tracks, the state-space model is updated with the assigned measurements for each track. The current state of the tracks are estimated using Kalman Filtering algorithm. The Kalman Filter algorithm and the used state-space model are explained in the following subsections.

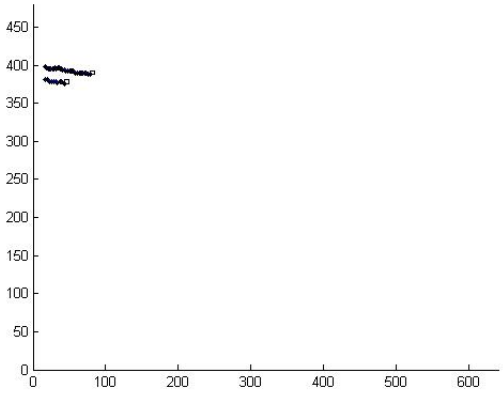
5.4.1 Kalman Filtering

Kalman Filter is first introduced by Rudolph E. Kalman in his famous paper [37]. It is used in many engineering fields because it can remove the noise while retaining the useful information and estimates the state of a linear dynamic system. A linear dynamic system is simply a process that can be described by the Equation 5.6 and 5.7 [36]:

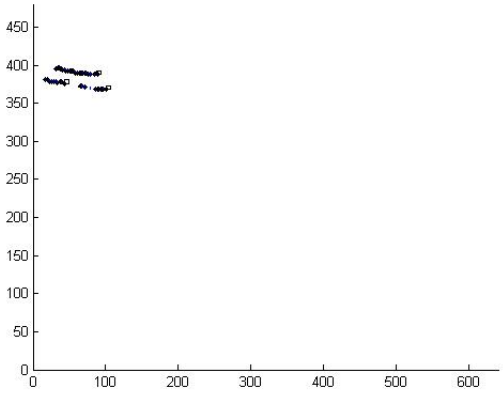
$$\text{State Equation:} \qquad x_{k+1} = Ax_k + Bu_k + w_k \tag{5.6}$$



(a) Frame 74



(b) Frame 85



(c) Frame 101

Figure 5.4: The output of Multi-Scan MCMCDA algorithm

Output Equation: $$z_k = Cx_k + v_k \tag{5.7}$$

where

- x_k is the state vector on time step k,
- z_k is the measurement on time step k
- A is the transition matrix that models the represents the dynamics of the models
- C_k is the measurement model since all of state may not be observed
- w_k and v_k are the process noise and measurement noises respectively

Kalman filter algorithm proposes a recursive solution for estimating the state of the linear dynamic system by using available noisy measurements while minimizing the variance of the estimation error [36]. This filtering algorithm has two steps : the prediction step and update(correction) step (depicted in Figure 5.5). In the prediction step the next state of the system is predicted by using the previous measurements and in update (correction) step the current state is estimated by using the current available noisy measurement.

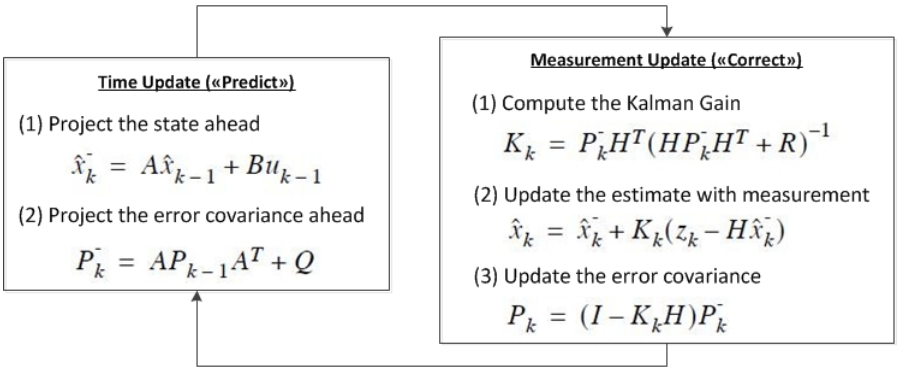


Figure 5.5: Kalman Filtering steps [36]

5.4.2 State Space Model

The head or upper body locations are measured in two dimensions in the video. So the state space used for modelling the dynamics of the surveillance world has two axis: x and y representing a cartesian coordinate system depicted in Figure 5.6.

Figure 5.6: Cartesian coordinate system for the video frames

A linear model is preferred for simplicity so the accelaration values are ignored for both axes. The velocity values in the cartesian coordinate system with cartesian locations form the state vector x_k at time step k.

$$\begin{bmatrix} x_k \\ y_k \\ V_{x_k} \\ V_{y_k} \end{bmatrix}$$

The transition matrix used for the linear model is defined as

$$\begin{bmatrix} 1 & 0 & \delta_t & 0 \\ 0 & 1 & 0 & \delta_t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

where δ_t is the time between two frames. For 25 fps videos this value is 0.04 sec.

The measurement model is defined as

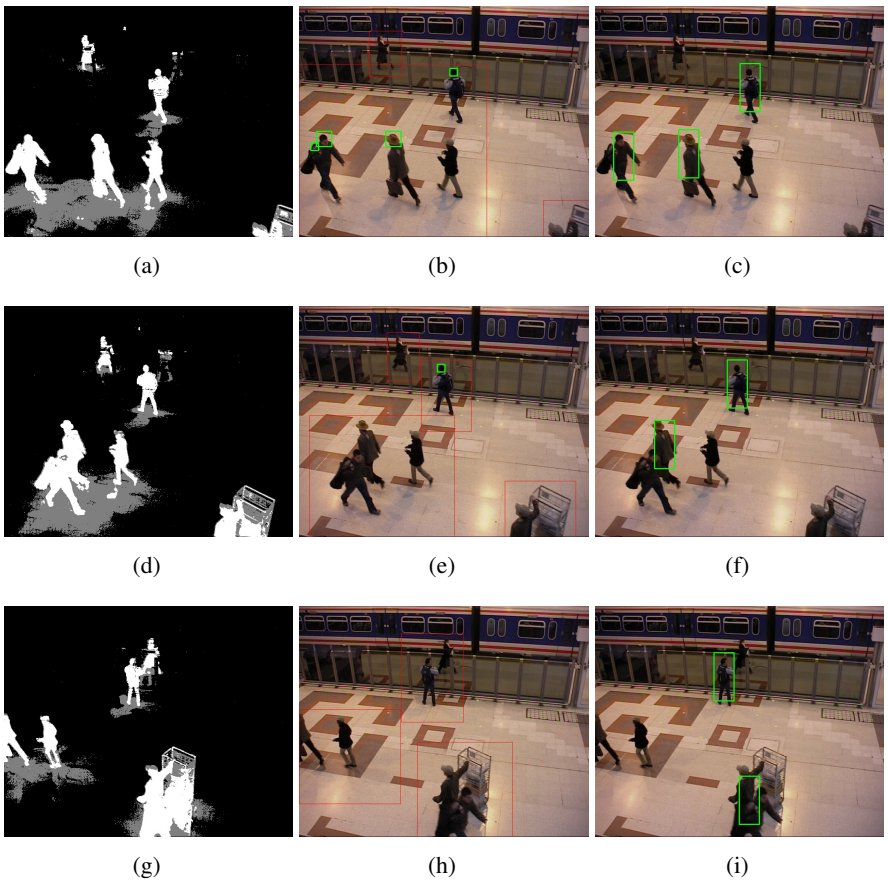
$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

In videos, the measurements include only the location of objects; however their velocities are unknown. The trained HOG detector (explained in Chapter 4) finds the head locations or upper body locations of objects of interest. So the chosen measurement model matrix provides only the locations of objects to solution domain.

5.5 Tracking Results

According to [38], there is no consensus on how to evaluate a tracker, and numerical evaluations are rare. Regarding the tracker as a detector, testing its accuracy at detection is seen as a fair evaluation by [38]. Since the detection mechanism is integrated to the tracking problem in this thesis, an extra evaluation has not been made. The detection results were explained in Chapter 4.

The multiple human tracking with human head detections and human upper body detections are depicted in Figure 5.7 and Figure 5.8. The tracker is tested over some scenarios using PETS2006 [39] video dataset. In Figure 5.7, (a)(d)(g) shows the foreground objects, (b)(e)(h) shows the head detections in pre-computed ROIs and (c)(f)(i) shows the estimated positions of the tracks. Although (e)(h) does not involve enough measurements, (f)(i) shows the positions which are estimated by Kalman Filtering algorithm. In Figure 5.8, (a)(d)(g)(j)(m) shows the foreground objects, (b)(e)(h)(k)(n) shows the human upper body detections in pre-computed ROIs and (c)(f)(i)(l)(o) the estimated positions of the tracks. Although (k)(n) do not involve enough measurements, (l)(o) show the positions which are estimated by Kalman Filtering algorithm.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 5.7: Head Tracking Results

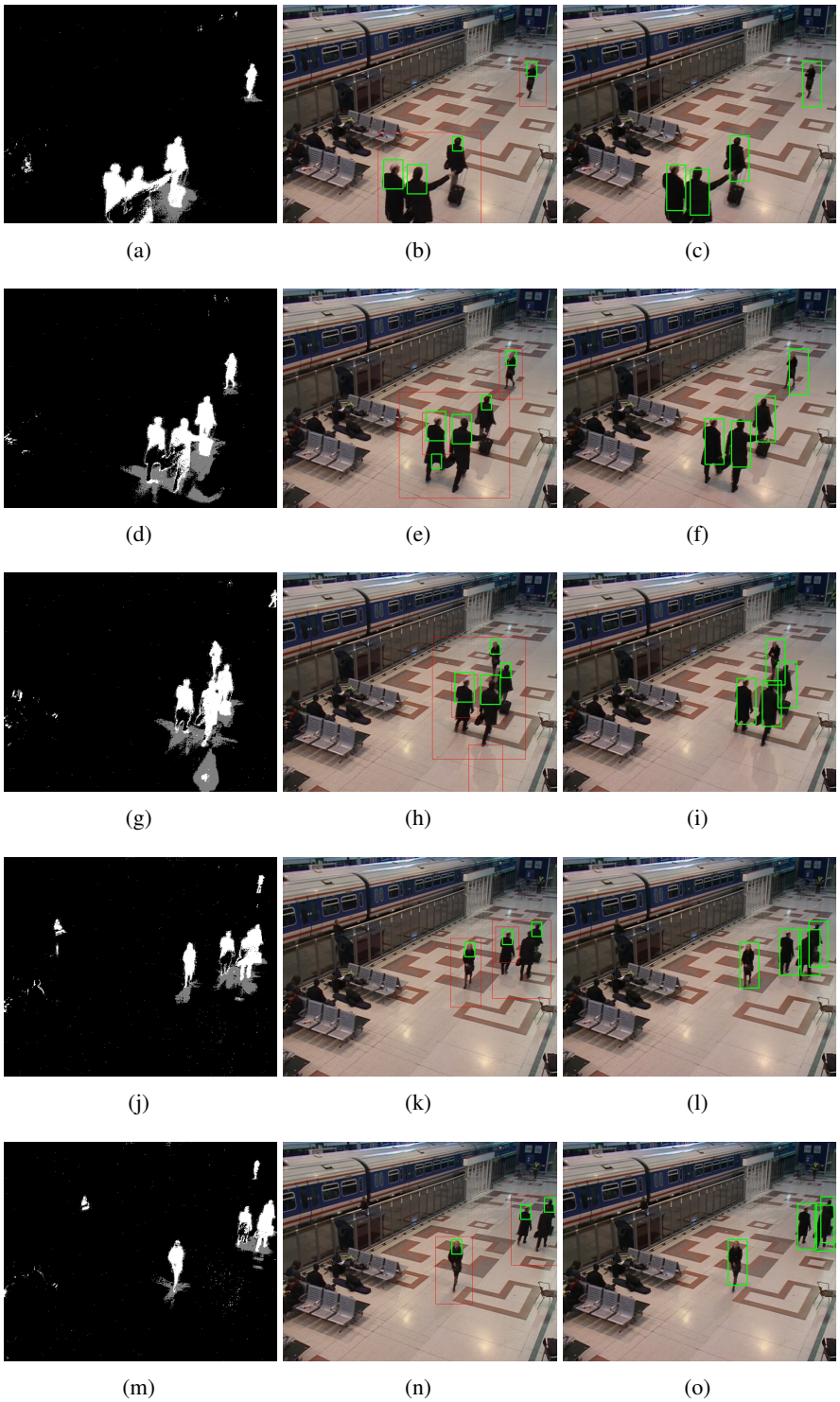
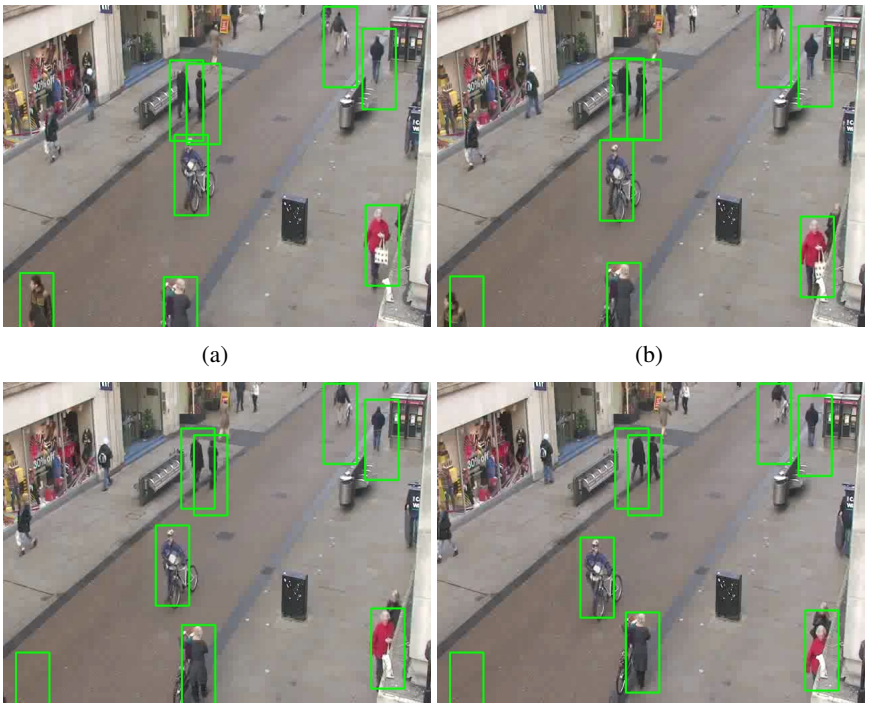


Figure 5.8: Upper Body Tracking Results

The detections of humans with pushing or carrying some objects are depicted in Figure 5.9. Since the detector finds the upper body part within the ROIs, the unrelated objects are ignored although they are occluded or stuck together. The tracking process is shown in Figure 5.10.

Figure 5.9: Partially occluded humans



(a) (b)

(c) (d)

Figure 5.10: Humans with objects

CHAPTER 6

CONCLUSION

6.1 Summary

This thesis concentrates on visual surveillance videos obtained from the crowded scenes such as train stations, streets and airports. A human detection based multi-human tracker is implemented. The starting point of this thesis is the concept of the integration of human detectors' advantages and robustness of the trackers. Human detectors can find humans accurately in the crowded scenes and trackers can track objects even in the cluttered environments.

The overall system has three main parts: Motion Detection, Human Detection and Multi-Human tracking. Each part is dependent on the former part. First, moving object regions are detected in Motion Detection part since these regions involve potential targets to be tracked. The boundaries with extra margins constitute the Region of Interests (ROIs) for human detection. In human detection part, the trained HOG descriptors are used for human head detection and upper body part detection. These detections constitute location measurements for object tracking.

The background is modeled by a Mixture of Gaussians (MOG) which is adaptive to dynamic environments for static cameras. When a new frame is captured, the foreground objects are extracted with some noise. The noise is removed by some morphological operations. Erosion followed by dilation are applied to the foreground image to remove noises. After the removal of noises, the connected component analysis method is applied to label each component. Since each component has a potential to have human(s) inside, ROIs for human detections are constituted from the components by adding extra margin to the boundaries of components.

In human detection part, human head detector and upper body part detector are seperately

trained using the Histogram of Oriented Gradients (HOG) descriptor. The training procedure is examined observing the effects of the HOG parameters in detail. The window size, the number of orientation bins (spaced over 0 - 180 degree) and the block overlapping rate are the HOG parameters that the experimentally observed in the training. In addition, the effects of gamma correction and hard training are observed. The head dataset of Oxford University - Active Vision Group is used for human head training and MIT Pedestrian dataset is used for human upper body training. The head training dataset consists of 1836 images with 24x24 pixels and the upper body training dataset consists of 934 image patches of size 32x32 pixel and 32x48 pixel. The number of positive examples in these datasets are doubled by generating new images consisting the reflection of each image.

In both training phases high overlapping rates and low number of orientations bins give better recall performance. The application of gamma normalization does not affect the recall rate for human head detection and human upper body part detection. Finally, the human head detector and upper body detector are re-trained in order to reduce the false positive rates, called Hard Training. Although the reduction in false positive rate is achieved, the false negatives are also increased. Hence the final human head detector that generates measurements for the tracker is formed with 24x24 window size, 9 orientation bins, 75% block overlapping rate. On the other hand, the final human upper body detector is formed with 32x48 window size, 9 orientation bins and 75% block overlapping rate. The training processes of final detectors do not involve the hard training phase due to the high rate of false negatives.

The human upper body detector generates less false positives compared to the human head detector because the omega-like shape of the human upper body is more descriptive to the shape of human head. Since the window size of human head is smaller than the human upper body, the duration of finding the human heads are smaller than human upper body in the same scene.

After the detections generate the location measurements, these measurements are associated to current tracks using Multi Scan Markov Chain Monte Carlo Data Association Algorithm (MCMCDA). In Multi-Scan MCMCDA, the recent history of measurements are utilized for computing the posteriori instead of measurements belongs to only current frame. Due to its simplicity, this algorithm is suitable for real time multi target tracking compared to MHT and JPDAF. The associated measurements to tracks are used for the update of their states. The

states of the tracks are estimated using Kalman Filtering algorithm. 4-dimensional state vector consisting the locations at two axes x and y and the speeds at two axes is used for state-space modelling.

The overall system has the capabilities of detecting and tracking multi-humans in a scene captured from a static camera. Although no real-time requirements are included in the scope of the thesis, Multi-Scan MCMCDA algorithm meets the real time requirements handling the newly births of humans into the scene and deaths from the scene. Apart from human tracking, the human detector itself has the ability of finding humans in the images. So this capability can be used in different domains such as content based image/video retrievals, automatic/semi-automatic semantic annotation tools.

The implemented multi-human tracker system can be improved in terms of the accuracy of the detections, speed and robustness. The future works are listed in the following section.

6.2 Future Work

In this thesis, the advantages and disadvantages of human head detector and human upper body detector are discussed. The hybrid version of these detectors can be a remedy for false negative and false positive rates. Different configurations of head and upper body locations can be fused with data fusion algorithms in order to increase the accuracy. This improvement will directly affect the robustness of the multi-human tracker. On the other hand, an adaptive thresholding mechanism can be applied for the detection decision. The trained detectors use a threshold value in order to decide if the detection window contains a human or not. This threshold value can be dynamically changed (probably reducing the threshold) within an estimated region of interest. This estimated region of interest can be computed by Kalman Filter.

The camera and processing unit technologies are being improved from day to day. In the future, we will have higher resolution images and videos ever, and processing of these images and videos gain more importance especially for biometric analysis applications that require identification of targets. The speed of the tracking system should meet the real time requirements for real world applications. At this point, the parallelism of the algorithms targetting the GPU (Graphical Processing Unit) environments is one of the solution. So moving the

current implemented system into a GPU environment will be another future direction.

The current developed system is suitable for static cameras. This study can be extended for active cameras having the pan-tilt-zoom capabilities for real world problems. With this extension, the trained detectors will not be affected but the background modelling system will sense that all pixels are moving when the camera is panned or tilted. In this situation, the region of interest for human detection will be the whole captured scene that would cause performance problems. If the current system is implemented for GPU environments, this performance problem can be ignored.

REFERENCES

[1] H. Dee and S. Velastin, *How close are we to solving the problem of automated visual surveillance? Machine Vision and Applications*, 19(5-6):329-343, 2008.

[2] B. A. Nabil. *Smart Cameras*, Springer, 2010.

[3] *Mind's Eye*, http://www.darpa.mil/Our_Work/I2O/Programs/Minds_Eye.aspx, last accessed date : 19.03.2012.

[4] W. Hu, T. Tan, L. Wang and S. Maybank, *A Survey Of Survaillance of object motions and behaivours*. IEEE Transactions on Systems, Man. and Cybernetics, vol. 34, no. 3, Aug 2004.

[5] I. S. Kim, H. S. Choi, K. M. Yi, J. Y. Choi, and S. G. Kong, *Intelligent Visual Surveillance - A Survey*. International Journal of Control, Automation, and Systems, 2010.

[6] P.Turaga, R. Chellappa, V. S. Subrahmanian and O. Udrea. *Machine recognition of human Activities : A Survey*. IEEE Trans. Circuits, Syst. Video Technol. vol. 8, no. 11, pp. 1473-1488, 2008.

[7] W. T. Freeman, *Where computer vision needs help from computer science*, ACM-SIAM Symposium on Discrete Algorithms (SODA), January, 2011, invited talk.

[8] B. Schiele, M. Andriluka, S. Roth and C. Wojek. *Visual People Detection - Different Models, Comparison and Discussion*. IEEE International Conference on Robotics and Automation, 2009.

[9] M. Andriluka, S. Roth and B. Schiele. *People-Tracking-by-Detection and People-Detection-by-Tracking*. IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[10] P. Felzenszwalb, D. McAllester and D. Ramanan. *A discriminatively trained, multiscale, deformable part model*. in IEEE Conference on Computer Vision and Pattern Recognition,2008.

[11] N. Dalal and B. Triggs.*Histogram of Oriented Gradients for human detection*. In IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 886-893, 2005.

[12] C. Papageorgiou and T. Poggio. *A trainable system for object detection*. International Journal of Computer Vision(IJCV),38(1):15-33, 2000.

[13] N. Dalal, *Finding People in Images and Videos*, Institut National Polytechnique de Grenoble, 2006.

[14] A. Yilmaz, O. Javed, and M. Shah. *Object Tracking: A Survey*. ACM Computing Surveys, vol.38, no.4,2006.

[15] S. Oh, S. Russell and S. Sastry. *Markov Chain Monte Carlo Data Association for Multiple-Target Tracking*. IEEE Transactions on Automatic Control, vol. 54, no. 3, pp. 481-497, Mar 2009.

[16] D.N. Ta, W.C. Chen, N. Genfald, K. Pulli, *SURFTrac : Efficient Tracking and Continous Object Recognition using Local Feature Descriptors*, CVPR 2009, pp. 2937-2944, June 2009.

[17] H. Zhou, Y. Yuan, C. Shi, *Object tracking using SIFT features and mean shift*, Computer Vision and Image Understanding, vol. 113, Issue 3, pp 345-352, March 2009.

[18] H. Liu, Ze Yu, H. Zha, Y. Zou, L. Zhang, *Robust human tracking based on multi-cue integration and mean-shift*, Pattern Recognition Letters, vol. 30, Issue 9, pp 827-837, July 2009.

[19] J.M. del Ricon, D. Makris, C.O. Urunuela, J.C. Nebel, *Tracking Human Position and Lower body parts using Kalman and Particle Filters constrained by Human Biomechanics*, IEEE Systems, Man and Cybernetics , Part B: Cybernetics, pp. 26-37 February 2011.

[20] J. Ashida, R. Miyamoto, H. Tsutsui, T. Onoye, Y. Nakamura, *Probabilistic Pedestrian Tracking based on a skeleton model*, IEEE International Conference on Image Processing, October 2006.

[21] T. Tung, T. Matsuyama, *Human Motion Tracking using a Color-Based Particle Filter Driven by Optical Flow*, MLVMA 2008.

[22] K.F. Sim, K. Sundaraj, *Human motion tracking of athlete using optical flow & artificial markers* , ICIAS 2010

[23] C. J. Needham , R. D. Boyle, *Tracking multiple sports players through occlusion, congestion and scale*, vol. 1, pp. 93-102 BMVA, 2001.

[24] B. Benfold, I. Reid, *Stable multi-target tracking in Real-time surveillance Video*. IEEE Conference on Computer Vision and Pattern Recognition, 2011.

[25] X. Ji and H. Liu, *Advances in View-Invariant Human Motion Analysis: A Review*, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 40, pp. 13-24, 2010.

[26] C. Stauffer and W.E.L. Grimson, *Learning Patterns of Activity Using Realtime Tracking*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 747-757, August 2000.

[27] T. B. Moeslund , A. Hilton and V. Krüger, *A survey of advances in vision-based human motion capture and analysis*, Computer Vision and Image Understanding, v.104 n.2, p.90-126, November 2006.

[28] S. Belongie, J. Malik, and J. Puzicha, *Shape Context: A new descriptor for shape matching and object recognition*, Neural Information Processing Systems, 2000.

[29] B. Leibe, E. Seemann and B. Schiele, *Pedestrian detection in crowded scenes*, CVPR, pp.878-885, 2005.

[30] A. Kläser and M. Marszałek and C. Schmid and A. Zisserman. *Human Focused Action Localization in Video*. International Workshop on Sign, Gesture, and Activity (SGA) in Conjunction with ECCV, 2010.

[31] S.S. Blackman. *Multiple Hyphothesis Tracking For Multiple Targets*. IEEE Aerospace and Electronic Systems, vol. 19, no. 1, Part 2 : Tutorials,Jan 2004.

[32] Y. Bar-Shalom, F. Daum and J. Huang. *The Probabilistic Data Association Filter*. IEEE Control Systems, vol. 29, pp.82-100, Nov 2009.

[33] H. Pasula, S. J. Russell, M. Ostland, and Y. Ritov, *Tracking many objects with many sensors*. International Joint Conference on Artificial Intelligence, Stockholm, 1999.

[34] *Convolution Kernel Mask Operation*, http://micro.magnet.fsu.edu/primer/java/digital imaging/processing/kernelmaskoperation/, last accessed date : 21.03.2012.

[35] *Metropolis-Hastings algorithm*, http://en.wikipedia.org/wiki/Metropolis%E2%80%93 Hastings_algorithm, last accessed date : 19.03.2012.

[36] G. Welch and G. Bishop, *An Introduction to Kalman Filter*, July 2006.

[37] R. E. Kalman, *A New Approach to Linear Filtering and Prediction Problems*, Transaction of the ASME-Journal of Basic Engineering, pp. 35-45, Mar 1960.

[38] D.A. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien and D. Ramanan, *Computational Studies of Human Motion : Part 1, Tracking and Motion Synthesis*, Foundations and Trends in Computer Graphics and Vision, vol.1, no. 2/3, pp. 77-254, 2006

[39] University of Reading, *PETS 2006*, http://ftp.pets.rdg.ac.uk/PETS2006/,September 2009.