CHAPTER 1

INTRODUCTION

1.1 PREAMBLE

Video surveillance in a dynamic scenario has a wide spectrum of promising applications such as access control in restricted areas, human identification in a crowd, crowd flux statistics and congestion analysis, detection of anomalous behaviours and interactive surveillance (Hu et al 2004). The strong need of smart surveillance systems stems from those security-sensitive locations such as banks, department stores, parking lots, hospitals and hotels.

The conventional procedure is that, the outputs of a surveillance camera is usually recorded in tapes or stored in video archives. These video data are used just as a forensic tool, losing its primary benefit as an active real-time media. The goal of video surveillance is not only to put cameras in place of human eyes, but also to accomplish automated surveillance task as much as possible. Hence, the need of the hour is to carryout real-time analysis of surveillance data to alert security system while an offence is underway. The issues of video surveillance in dynamic scenes are detection, classification, tracking and recognition of certain objects in an image sequence, and more generally understanding the behaviour of an object. Various computer vision based technologies have been proposed in the literature to address these challenges (Aggarwal and Cai 1999).
One of the pioneering systems reported for people tracking is Pfinder (‘Person Finder’), proposed by Wren et al (1997) and developed at MIT media labs, employs the maximum a posteriori (MAP) probability models to detect human body in 2D image planes, especially in indoor scenes. In Europe, Isard and Blake (1998) of Oxford University proposed conditional density propagation approach to track moving objects. The defence advanced research projection agency (DARPA) supported the visual surveillance and monitoring (VSAM) project in 1997, whose purpose was to develop an automatic video understanding technologies that enable a single human operator to monitor behaviours over complex areas such as battlefields and civilian scenes (Collins et al 2000). Furthermore, to enhance protection from terrorist attacks, DARPA sponsored a program for human identification at a distance (HID) in 2000 (Collins et al 2000). The famous W4 (‘What, Where, When, Who’) system developed at University of Maryland by Haritaoglu et al (2000) was, able to detect multiple people in the outdoors and to analyse body silhouette for inferring people’s activity. Comaniciu and Meer (2002) of Siemens labs, U.S.A, proposed the tracking algorithm based on mean-shift techniques to follow the human in crowded environments. Connell et al (2004) of IBM proposed appearance-based tracking in cluttered indoor environments.

1.2 VIDEO SURVEILLANCE SYSTEM

The processing framework of video surveillance in dynamic scenes includes the modeling of environments, motion segmentation, classification of moving objects, tracking, understanding and description of behaviours, human identification as shown in Figure 1.1. Motion segmentation is to extract the moving foreground objects from the background. Classification of the moving objects in natural scenes involves, providing the meaningful information to such foreground objects. Surveillance based moving object tracking over time typically involves matching objects in consecutive frames using features such
as points, centroid or bounding box and the moving objects are labeled concurrently. The image analysis and recognition of motion patterns are carried out to understand the behaviour and to produce high level description of actions and interactions (Hu et al 2004).

![Diagram of a general video surveillance system]

**Figure 1.1 A General Video Surveillance System**

### 1.2.1 Environment Modeling

Environment modeling refers to construction and representation of data pertaining to background. This can be classified into 2-D models in the image plane and 3-D models in real world coordinates. Also, the environment is categorized in accordance with the camera position such as fixed camera, pan tilt zoom (PTZ) and moving camera. For fixed cameras, the key problem is to automatically recover and update background images from a dynamic sequence. Unfavorable factors, such as illumination variance, shadows and
shaking branches, bring many difficulties to the acquisition and updation of background images.

There are many algorithms for resolving these problems including temporal average of an image sequence, adaptive Gaussian estimation, and parameter estimation based on pixel processes. Ridder et al (1995), modeled each pixel value with a Kalman Filter to compensate for illumination variance. Stauffer and Grimson (1999) presented a theoretic framework for recovering and updating background images based on a process in which a mixed Gaussian model was used for each pixel value and online estimation is used to update background images in order to adapt to illumination variance and disturbance in backgrounds. Toyama et al (1999) proposed the Wallflower algorithm in which background maintenance and background subtraction were carried out at three levels: the pixel level, the region level, and the frame level. Haritaoglu et al (2000) built a statistical model by representing each pixel with three values: its minimum, the maximum intensity values and the maximum intensity difference between consecutive frames observed during the training period. These three values were updated periodically in order to maintain the background model. McKenna et al (2000) used an adaptive background model with color and gradient information to reduce the influences of shadows and unreliable color cues.

Pan-tilt-zoom (PTZ) environment model considers the scene by viewing in all directions by changing the position of the camera. In such model, pure translation can be realised by patching up a panorama graph to acquire a holistic background image. Homography matrices are used to describe the transformation relationship between different images. In the case of moving cameras, motion compensation is needed to construct temporary background images. The research in 3-D environmental model is still limited
to indoor scenes because of the difficulty of 3-D reconstructions of outdoor scenes.

1.2.2 Motion Segmentation

Motion segmentation in image sequences aims at detecting regions corresponding to moving objects such as vehicles and humans. Detecting moving regions provides a focus of attention for later processes such as tracking and behaviour analysis because only these regions need to be considered. Most segmentation methods use either temporal or spatial information in the image sequence. The conventional approaches for motion segmentation are background subtraction, temporal differencing and optical flow.

1.2.2.1 Background subtraction

Background subtraction is a popular method for motion segmentation, especially under for images having static background. It detects moving regions in an image by taking the difference between the current image and the reference background image in a pixel-by-pixel analysis. It is simple, but extremely sensitive to changes in dynamic scenes derived from lighting and extraneous events such as camouflage, shadow. Hence, it requires robust background model to nullify the influence of these changes. Cucchiara et al (2003) presented two methods to model the background: First one is the simple pixel based background subtraction that operates on pixel level. The bottle neck is, it cannot distinguish specific objects in a scene, such as humans, cars and it is more likely to classify every new object it sees in the background as foreground object. This method is largely used as a pre-processing step in a vision system. The second method, namely model- based background subtraction is basically statistical method uses aprior knowledge about the object for further processing.
1.2.2.2 Temporal Differencing

Temporal differencing makes use of the pixel-wise differences between two or three consecutive frames in an image sequence to extract moving regions. It is very adaptive to dynamic environments, but generally does a poor job of extracting all the relevant pixels, and sometimes develops holes that results in error. Lipton et al (1998), detected moving targets in real video streams using temporal differencing. After the absolute difference between the current and the previous frame is obtained, a threshold function determines the changes. By using a connected component analysis, the extracted moving sections are clustered into motion regions.

1.2.2.3 Optical Flow

Optical-flow-based motion segmentation uses characteristics of flow vectors of moving objects over time to detect moving regions in an image sequence. Meyer et al (1997), computed the displacement vector field to initialize a contour based tracking algorithm, called active rays, for the extraction of articulated objects that are used for gait analysis. Optical-flow-based methods can be used to detect independently moving objects even in the presence of camera motion. However, most flow computation methods are complex, very sensitive to noise and cannot be applied to video streams in real time without specialized hardware. More detailed discussion of optical flow is listed in the work of Barron et al (1994).

Besides the basic methods described above, there are few other approaches for motion segmentation. Using the extended expectation maximization (EM) algorithm, Friedman and Russell (1997), implemented a mixed Gaussian classification model for each pixel. In their work, slowly moving objects are handled perfectly and shadows are eliminated effectively.

Amongst the methods, background subtraction is a widely used approach for detecting moving objects from static cameras. Piccardi et al (2004) provided a review of the main methods and categorize the methods based on speed, memory requirements and accuracy. The methods for background subtraction are running Gaussian average, temporal median filter, mixture of Gaussians, kernel density estimation (KDE), sequential kernel density approximation, co-occurrence of image variations and Eigen backgrounds. Amongst the methods, simple methods such as the running Gaussian average or the median filter offer acceptable accuracy while achieving a high frame rate and having limited memory requirements (Wren et al 1997). Methods such as Mixture of Gaussians (MoG) and kernel density estimation (KDE) prove very good model accuracy. But, kernel density estimation has a high memory requirement which might prevent easy implementation on low-memory devices. Sequential kernel density estimation is an approximation of KDE which proves almost as accurate, but mitigates the memory requirement by an order of magnitude and has lower time complexity (Elgammal et al 2000). Methods such as the co-occurrence of image variations and the Eigen backgrounds explicitly address spatial correlation. They both offer reasonable accuracy against time and memory complexity. However, practical implementation of the co-occurrence method imposes a trade off with resolution. In this scenario, it is evident that the Mixture of Gaussians (MoG) method surpasses other approaches in terms of speed, memory accuracy, complexity and shadow elimination. Since, the most
significant factor that affects the performance of background subtraction algorithm is the shadow of the object amongst illumination, clutter and camouflage.

Hence, the above methods may also be looked in the perspective of shadow elimination capability. Though various probabilistic and mixture models, parametric model and nonparametric models are available to cater to the issue of shadows, it is beneficial to use Mixture of Gaussians (Stauffer and Grimson 1999), to eliminate the shadow, since it is simple to implement and accurate enough to remove moving shadows.

1.2.3 Object Classification

The next step in video surveillance system is object classification. Different moving regions may correspond to different moving targets in natural scenes. For instance, the image sequences captured by surveillance cameras mounted in road traffic probably include humans, vehicles and other moving objects such as flying birds and moving clouds, where the object of interest is only human and vehicle. Hence it is essential to classify the required moving objects correctly for further processing. At present, there are two main categories of approaches for classifying moving objects are reported in the literature, namely shape based classification and motion based classification.
1.2.3.1 Shape-based Classification

Shape based classification depends on the descriptions of shape information of motion regions such as points, boxes and silhouettes for classifying moving objects. VSAM takes object’s dispersedness, area, apparent aspect ratio of bounding box as key features, and classifies moving-objects into four classes: single human, vehicle, group of people and clutter (Collins et al 2000). Kuno et al (1996) used simple shape parameters of human silhouette patterns to separate humans from other moving objects. Lipton et al (1998) used the dispersedness and area of objects as classification metrics to classify all moving-objects into human, vehicle and clutter. Temporal consistency constraints are considered in their analysis in order to obtain precious classification results.

1.2.3.2 Motion-based Classification

In general, non rigid articulated human motion shows a periodic property, so this has been used as a strong cue for classification of moving objects. Lipton (1999) used residual flow to analyse rigidity and periodicity of moving objects. It is shown that rigid objects present little residual flow, whereas a non rigid moving object such as a human being has a higher average residual flow and even display a periodic component. Based on this useful cue, human motion is distinguished from motion of other objects, such as vehicles. Cutler and Davis (2000) described a similarity-based technique to detect and analyse periodic motion. By tracking the moving object, its self-similarity is computed as it evolves over a period of time which is the measure of periodicity. Therefore time-frequency analysis is applied to detect and characterize the periodic motion.
The two common approaches mentioned above, namely shape-based and motion-based classification can also be effectively combined for classification of moving objects. Furthermore, Stauffer and Grimson (2000) proposed a novel method based on a time co-occurrence matrix to hierarchically classify both objects and behaviours. It is expected that more precise classification results can be obtained by using extra features such as color and velocity.

1.2.4 Object Tracking and Labeling

After motion detection, surveillance systems generally track moving objects from one frame to another in an image sequence. Tracking over time typically involves matching objects in consecutive frames using features such as points, lines or blobs. Useful mathematical tools for tracking include the Kalman filter, the condensation algorithm, the dynamic Bayesian network and the geodesic method. Tracking methods are divided into four major categories: region-based tracking, active-contour-based tracking, feature based tracking, and model-based tracking. Moeslund et al (2006) pointed out in his survey that this classification is not absolute but algorithms from different categories can be integrated together. After reliable tracking of multiple people consistent labeling is to be done to facilitate further processes successfully. Probabilistic based approaches and stochastic sampling have been introduced to improve the reliability of tracking and labeling (identification of moving objects) during occlusion.

1.2.4.1 Region-Based Tracking Algorithms

Region-based tracking algorithms track objects according to variations of the image regions corresponding to the moving objects. Wren et al (1997) explored the use of small blob features to track a single human in an
indoor environment. In their work, a human body is considered as a combination of some blobs respectively representing various body parts such as head, torso and the four limbs and by tracking each blob, the moving human is successfully tracked. McKenna et al (2000) proposed tracking algorithms at three levels of abstraction: regions, people, and groups. Each region has a bounding box and regions can merge and split and considers a human is composed of one or more regions grouped together. Therefore, using the region tracker and the individual color appearance model, perfect tracking of multiple people is achieved. Although they work well in scenes containing only a few objects, region-based tracking algorithms cannot reliably handle occlusion between objects (Hu et al 2004). Accordingly, these algorithms cannot satisfy the requirements for surveillance against a cluttered background or with multiple moving objects.

1.2.4.2 Active Contour-Based Tracking Algorithms

Active contour-based tracking algorithms track objects by representing their outlines as bounding contours and updating these contours dynamically in successive frames. These algorithms aim at directly extracting shapes of subjects and provide more effective descriptions of objects. Isard and Blake (1996) adopted stochastic differential equations to describe complex motion models, and combine this approach with deformable templates to cope with people tracking. Bregler and Malik (1998) have successfully applied active contour-based methods to vehicle tracking. Paragios and Deriche (2000) detected and tracked multiple moving objects in image sequences using a geodesic active contour objective function and a level set formulation scheme. Peterfreund (2000) explored a new active contour model based on a Kalman filter for tracking non rigid moving targets such as people in spatio-velocity space. In contrast to region-based tracking algorithms, active contour-based algorithms describe objects more simply and
more effectively and reduce computational complexity. Even under
disturbance or partial occlusion, these algorithms may track objects
continuously. However, the tracking precision is limited at the contour level.
A further difficulty is that the active contour-based algorithms are highly
sensitive to the initialization of tracking automatically (Wu and Huang 2001).

1.2.4.3 Feature-Based Tracking Algorithms

Feature-based tracking algorithms perform tracking of objects by
extracting elements, clustering them into higher level features and then
matching the features between images. Feature-based tracking algorithms can
further be classified into three subcategories according to the nature of
selected features: global feature-based, local feature-based and dependence-
graph-based algorithms. The features used in global feature-based algorithms
include centroids, perimeters, areas, some orders of quadratures and colors.
Polana and Nelson (1994) provided a good example of global feature-based
tracking. A person is bounded with a rectangular box whose centroid is
selected as the feature for tracking. Even when occlusion happens between
two persons during tracking, as long as the velocity of the centroids can be
distinguished effectively, tracking is still successful. The local feature based
algorithms and dependence graph based algorithms include line segments,
curve segments, and corner vertices (Coifman et al 1998). However, there are
several serious deficiencies in feature-based tracking algorithms. The
recognition rate of objects based on 2-D image features is low, because of the
nonlinear distortion during perspective projection and the image variations
with the viewpoint’s movement. The stability of dealing effectively with
occlusion, overlapping and interference of unrelated structures is generally
poor (Schiele 2000).
1.2.4.4 Model-Based Tracking Algorithms

Model-based tracking algorithms track the objects by matching projected object models, produced with prior knowledge, to image data. The models are usually constructed off-line with manual measurement, CAD tools and computer vision techniques (Bregler 1997). Focused on model-based human body tracking, the general approach is known as analysis-by-synthesis, and it is used in a predict-match-update style. Firstly, the model of the pose for the next frame is predicted according to prior knowledge and tracking history. Then, the predicted model is synthesized and projected into the image plane for comparison with the image data. A specific pose evaluation function is needed to measure the similarity between the projected model and the image data (Zhao et al 2002). Generally, model-based human body tracking involves two main issues: construction and representation of human models.

a) Construction of Human body models:

Construction of human body models is the base for model-based human body tracking. The accuracy of the tracking depends on the complexity of the human body model that results in expensive computation. Traditionally, the geometric structure of human body can be represented in the following four categories such as stick figure, 2-D contour, volumetric models and 3-D models.

The essence of human motion is typically contained in the movements of the torso, the head and the four limbs. In the stick-figure, the parts of a human body is represented as sticks and the sticks are linked with joints. Karaulova et al (2000) used a stick figure representation to build a hierarchical model of human dynamics encoded using Hidden Markov Models (HMMs), and realize view-independent tracking of a human body in monocular image sequences. In the 2-D contour representation, the human
body model is directly represented by human body projections in an image plane. The human body segments are modeled by 2-D ribbons. Ju et al (1996) proposed a cardboard human body model, in which the human limbs are represented by a set of jointed planar ribbons. The main disadvantage of 2-D models is that they require restrictions on the viewing angle. To overcome this disadvantage, volumetric models are proposed but they require more parameters than image-based models and lead to more expensive computation during the matching process. Wachter and Nagel (1999) established a 3-D body model using connected elliptical cones. In hierarchical model, Plankers and Fua (2001) presented a hierarchical human model having skeleton. This model includes four levels such as skeleton, ellipsoid meatballs simulating tissues and fats, polygonal surface representing skin and shaded rendering to get accurate tracking results.

b) Representation of Motion models:

Motion models of human limbs and joints are widely used in tracking. They are effective because the movements of the limbs are strongly constrained. These motion models serve as prior knowledge to predict motion parameters, to interpret and recognize human behaviours, or to constrain the estimation of low-level image measurements. Ning et al (2002) presented a motion model from semi-automatically acquired training examples and represent it using Gaussian distributions.

Tracking objects is equivalent to establishing the correspondence of image features between frames. Though the following four approaches based on model, active-contour, region and feature are available for the task of tracking, human activity analysis in image sequences assumes that feature based tracking is suitable. Because, the features such as appearance, shape and motion have increased the reliability of tracking people with partial occlusion.
Still, for specific environments, there remains a gap between the state-of-the-art tracking and consistent labeling of people.

1.2.5 Person-Specific Identification

In vision-based human identification at a distance, gait is a most attractive modality. Generally, gait recognition will focus on the following directions. Gait includes both individual appearances and the dynamics of walking. Developing the underlying static parameters of a human body and the dynamic characteristics of joint angles is helpful for recognition. There are four categories of person specific identification methods such as Model based methods, statistical methods, physical parameter based methods and spatio-temporal based methods.

1.2.5.1 Model-Based Methods

In model-based methods, parameters such as joint trajectories, limb lengths, and angular speeds, are measured from human silhouette. Cunado et al (1997) modeled gait as the movement of an articulated pendulum and use the dynamic Hough transform to extract the lines representing the thigh in each frame. Yam et al (2001) proposed a new biomechanical model-based gait recognition algorithm. Models of walking and running are used to form a type of new anatomical model called a dynamically coupled oscillator, for the hip motion, and the structure and motion of the thigh and the lower leg. However, accurately recovering features using model from a walking video is still an unsolved or not well-solved problem. In addition, the computational cost of the model-based approaches is quite high due to the complexity involved.
1.2.5.2 Statistical Methods

Statistical recognition techniques usually characterize the statistical description of motion image sets, and have been well developed in automatic gait recognition. Lee and Grimson (2002) used the moment features of image regions to recognize the individuals. Assuming that people walk frontal parallel toward a fixed camera, the silhouette region is divided into seven sub-regions. A set of moment-based region features is used to recognize people and to predict the gender of an unknown person by his/her walking appearance. Statistical methods are relatively robust to noise and change of time interval in input image sequences. Compared with model-based approaches, the computational cost of statistical methods is low.

1.2.5.3 Physical-Parameter-Based Methods

Physical-parameter-based methods make use of geometric structural properties of a human body to characterize a person’s gait pattern. The parameters used include height, weight, stride cadence and length. For example, the method proposed by Bobick and Johnson (2001) did not directly analyse the dynamics of gait patterns but uses walking activities to recover the static body parameters of walking such as the distance between body parts. However, they depend greatly on the vision techniques that are used to recover the required parameters like body-part labeling, depth compensation and camera calibration. In addition, the parameters used for recognition may be not effective across a large population.

1.2.5.4 Spatio-Temporal Motion-Based Methods

Niyogi and Adelson (1994) used the spatio-temporal surface to analyse gait. After motion detection, the spatio-temporal pattern (2-D space and 1-D time) is fitted with a smooth spatio-temporal surface. Spatio-temporal
motion-based methods are able to capture both spatial and temporal information of gait motion. Their advantage is low computational complexity and a simple implementation. However, they are susceptible to noise and to variations of the timings of movements.

The issues in gender classification are statistical learning of gait appearance, context-dependent learning from example images, handling occlusion while objects are moving, availability of small databases, real-time performance required by gesture analysis and the selection of features for classification tools.

1.2.6 Understanding and Description of Behaviours

After successfully tracking and labeling the moving objects from one frame to another in an image sequence, the problem of understanding object behaviours from image sequences follows naturally. Behaviour understanding involves the recognition and analysis of motion patterns, and the production of high-level description of actions and interactions. It may simply be thought as the classification of time varying feature data that is matching an unknown test sequence with a group of labelled reference sequences representing typical behaviours (Bobick and Davis 2001). It is then obvious that a fundamental problem of behaviour understanding is to learn the reference behaviour sequences from training samples, and to devise both training and matching methods for coping effectively with small variations of the feature data within each class of motion patterns.

For video surveillance systems, it is necessary to analyse the behaviours of people and determine whether these behaviours are normal or abnormal. Efforts on these lines have been made is based on learned patterns of behaviours and anomalies (Wang et al 2003). When a detected behaviour
does not match the learned patterns, it is classified as an anomaly. The object
behaviour can be predicted by matching the observed sub-behaviour of the
object with the learned patterns. Generally, patterns of behaviours in a scene
can be constructed by supervised or unsupervised learning of each object’s
velocities and trajectories (Fernyhough et al 2000). Supervised learning is
used for known scenes where objects move in pre-defined ways. For unknown
scenes, patterns of behaviours should be constructed by self-organizing and
self-learning of image sequences. Unsupervised learning is used to generate
action classes out of a large training set. These action classes are then used to
label test images. These approaches use a technique for deformable matching
of edges of image pairs, based on linear programming relaxation techniques
(Wang et al 2006). The major existing methods for behaviour understanding
are outlined in the following.

1.2.6.1 Finite State Machine (FSM)

The most important feature of a FSM is its state-transition function.
The states are used to decide which reference sequence matches with the test
sequence. Bobick and Wilson (1995) analysed the explicit structure of natural
gestures where the structure is implemented by an equivalent of a FSM but
with no learning involved. State-machine representations of behaviours have
also been employed in higher level description in defence applications.

1.2.6.2 Hidden Markov Models (HMMs)

A HMM is a kind of stochastic state machines. It allows a more
sophisticated analysis of data with spatio-temporal variability. The HMM
consists of two stages namely training and classification. In the training stage,
the number of states of HMM must be specified, and the corresponding state
transition and output probabilities are optimized in order that the generated
symbols can correspond to the observed image features. In the matching stage, the probability with which a particular HMM generates the test symbol sequence corresponding to the observed image features is computed. HMMs generally outperform other techniques and are therefore extensively applied to behaviour understanding. Starner et al (1998) proposed HMMs for the recognition of sign language. Oliver et al (2000) proposed and compared two different state-based learning architectures, namely, HMMs and Coupled Hidden Markov Models (CHMMs) for modeling people behaviour and interactions. Brand and Kettnaker (2000) showed that, by the use of the entropy of the joint distribution to learn the HMM, a HMM’s internal state machine can be made to organize observed behaviours into meaningful states. This technique has found applications in video monitoring and annotation, in low bit-rate coding of scene behaviours, and in anomaly detection.

1.2.6.3 Time Delay Neural Network (TDNN)

TDNN is also an interesting approach to analyse time-varying data. In TDNN, delay units are added to a general static network, and some of the preceding values in a time-varying sequence are used to predict the next value. As larger data sets become available, more emphasis is being placed on neural networks for representing temporal information (Yang and Ahuja 1998). TDNN has been successfully applied to hand gesture recognition and lip-reading (Meier et al 2000).

1.2.6.4 Syntactic Techniques

The syntactic approach in machine vision has been studied mostly in the context of pattern recognition in static images. Ivanov and Boblic (2000) describe a probabilistic syntactic approach to the detection and recognition of temporally extended behaviours and interactions between
multiple agents. The fundamental idea is to divide the recognition problem into two levels. The lower level is performed using standard independent probabilistic temporal behaviour detectors, such as HMMs, to extract possible low-level temporal features. These features are the input stream for a stochastic context-free parser. The higher level parser provides longer range temporal constraints, disambiguate uncertain low-level detection, and allow the inclusion of a priori knowledge about the structure of temporal behaviours in a given domain.

1.2.6.5 Non-Deterministic Finite Automation (NFA)

Wada and Matsuyama (2000) employed NFA as a sequence analyser for multi object behaviour recognition based on behaviour driven selective attention.

1.2.6.6 Self-Organizing Neural Network

All the methods discussed above are supervised learning and are applicable for known scenes where the types of object motions are known already. The self-organizing neural networks are suited to behaviour understanding when the object motions are unrestricted. Owens and Hunter (2000) applied the Kohonen self-organizing feature map to find the flow vector distribution patterns. These patterns are used to determine whether a point on a trajectory is normal or abnormal.

Recently, progress has been made towards action recognition in building the statistical models of human actions using machine learning. But for action classification, some restrictions are usually imposed to decrease ambiguity during matching of feature sequences (Schuld et al 2004). Therefore, the difficulties of action classification still lie in feature selection
and machine learning. Nowadays, the approaches of state space and template matching for action classification often choose a trade-off between computational cost and classification accuracy, so efforts should be made to improve performance of action classification, and at the same time to effectively reduce computational complexity (Wang and Li 2009). These ambiguities are exacerbated when several people are present in a scene, for example in a crowded environment.

1.3 CHALLENGES IN VIDEO SURVEILLANCE

The images captured in a dynamic environment are often affected by factors such as occlusion, shadow, clutter, lighting, and weather. Among these factors, the effect of shadows on the images in a dynamic environment, the problem of self-occlusion of human body and mutual occlusions between objects, consistent labeling of multiple people under congested conditions, the tracking of moving humans from one frame to another in an image sequence, the problem of gender identification, behaviour understanding by recognizing action are the issues being addressed by researchers around the world. The development of reliable, fast and accurate models should be adaptive to dynamic changes in complex environments is still a challenge.

1.4 MOTIVATION FOR THE PRESENT RESEARCH

The initiatives of global community to combat terrorism and to make the world a place for peaceful living has driven the computer vision community across the world to carry out active research in the area of video surveillance. Over five-hundred publications in the area of vision based human motion estimation have been made during the period 1997-2010. Inspired by these initiatives, the author chose to work in the topic of ‘computer
vision based algorithms on human action analysis for video surveillance’ so as to be part of global video surveillance research initiatives.

1.5 SCOPE OF THE THESIS

Computer Vision based human motion analysis has attracted great interests in significant areas such as video surveillance, video conferencing, athletic performance analysis, content-based image retrieval and virtual reality. This thesis is organized in a hierarchical manner (from low-level vision, intermediate-level vision, to high-level vision) according to a general framework of human motion analysis systems. The taxonomy selected is based on the functionalities including motion segmentation and object classification. Even though, many consolidated techniques have been tested for video surveillance applications using single fixed camera, distributed cameras, multi-modal acquisition such as pan-tilt-zoom (PTZ) cameras and moving cameras, in this thesis, fixed camera environment is considered. Three types of techniques such as background subtraction, temporal differencing and optical flow are addressed for motion segmentation. The background subtraction is chosen as a better choice for fixed camera environments.

Object classification is considered as a standard pattern recognition issue and shape and motion based approaches are two main categories for classifying moving objects. Non-rigid articulated human motion shows a periodic property, so motion based approaches has been used for classification of moving objects. Tracking and consistent labeling as an intermediate level is equivalent to establish correspondence of image features between frames. Four approaches such as model-based, active-contour-based, region-based and feature-based are conventional. Feature based tracking is considered in this thesis for the human action analysis. Semantic description of behaviours like gender classification and action classification are presented for the high-level
vision issues. Although a large amount of work has been reported in this area, many issues such as shadow removal, occlusion handling, consistent labeling during tracking, feature selection, trade-off between accuracy, speed and computational cost, personal identification, anomaly detection and behaviour prediction, content-based retrieval of surveillance videos, fusion of data from multiple cameras still remains unexplored.

Among these issues, this thesis addresses shadow removal for motion segmentation, occlusion handling for object classification, consistent labeling in a homogenous environment, gait based gender classification and action classification. The algorithms have been developed to overcome these issues and also provided the performance improvement of the surveillance system. The algorithms are evaluated quantitatively by detection rate, classification accuracy with reduced computational complexity using benchmark datasets.

1.6 ORGANISATION OF THE THESIS

Chapter 2 presents the effective foreground detection by eliminating shadow. The shadow is eliminated by modeling Gabor responses of the shadow pixels. Chapter 3 describes the moving object classification in the scene as human or vehicle using recurrent motion image analysis based on skeleton features. Chapter 4 discusses the consistent labeling of the human identification using skin color model in the homogeneous environment. Further, Chapter 5 presents the classification of the gender using gait appearance features for person specific identification. Chapter 6 describes the human action as ‘normal’ or ‘abnormal’ through projection and star skeleton features based on Relevance Vector Machine. A summary of the conclusions drawn from the research towards this thesis work is presented in Chapter 7.