Chapter 2

2 Background and Related work

Object tracking can be defined as the process of segmenting an object of interest from a video scene and keeping track of its motion, orientation, occlusion etc. in order to extract useful information. Visual tracking of the objects attempts to detect, track and identify the people or vehicles and interpret the object behavior from image sequences involving the objects.

In visual tracking two different approaches are merged together: Designing of Classifier and Motion Analysis of moving object. Different approaches used to classify the object are reviewed in the first section of this chapter. Different approaches used for designing the modules of object classifier like feature extraction, distance measures and performance metrics are reviewed.

In the second section, different algorithms and techniques used for visual tracking have been discussed. Different motion segmentation approaches using different background model and tracking methods have been reviewed and the merits and demerits of the each method have been discussed. From the different methods and reviews, the best approaches in terms of results and execution speeds are combined to make the efficient algorithms for object classification and visual tracking task.

2.1 Designing of Classifier

Designing of Classifier task involves different modules like feature extraction and distance measures. To validate the effectiveness of classifier, different performance metrics are measured. This section covers the literature review on the related work done so far in the above area.

2.1.1 Feature Extraction

A pattern recognition system that adjusts its parameters to find correct decision boundaries, through a learning algorithm using a training set such that a cost function (mean square error between numerically encoded values of actual and predicted labels) is minimized can be referred as a classifier or model [6].

Object Classification task uses two types of learning method: Supervised and Unsupervised learning. Supervised learning is the term used to describe the training of a classifier with target data available for the training set. The aim of object classifier is to find the correct mapping between input data and the target data. Unsupervised learning does not use target data. The goals of learning are finding clusters in data or modeling distributions as opposed to find a mapping.

Feature Extraction (A set of variables which carries discriminating and characterizing information about an object) and Feature Selection algorithms are mainly important for object classification. There are basically three approaches used for feature extraction [2], [4], [6]:

1. Geometry based approaches

2. Feature Point based approaches

3. Appearance based approaches

1. Geometry based approaches: Geometrical model based feature extraction can be done by extracting the geometric primitives like lines, curves or circles. They cannot handle the variation in the lighting and view points with certain occlusions. An excellent review on geometry based object recognition has been discussed in Mundy [10]. This paper reviews the key advances of the geometric era and the underlying causes of the movement away from formal geometry and prior models towards the use of statistical learning methods based on appearance features. Although geometry based approaches are invariants to view points and illumination, dependency and complexity on statistical functions have made limited use of the method.

2. Feature Point based approaches: The main idea of feature point based object recognition algorithm lies in finding interest points, often occurring at intensity discontinuities that are invariant to change due to scale, illumination and affine transformation. Feature Point based approaches find the different points that are invariant to the affine, rotation, translation or scaling. Various Feature Based Algorithms are reviewed such as Harris Corner Detector (HCD) [11], Scale Invariable Feature Transform (SIFT) [12], Speed up Robust feature Transform (SURF) [13], Random Sample Consenus (RANSAC) [14] etc.

Harris Corner Detector (HCD) method uses a combined corner and edge detector method based on the local correlation function to find out the image regions containing texture and isolated features. It shows good consistency and performance over a natural image. Scale Invariant Feature transform are invariant to image scaling, translation, rotation and partially invariant to illumination changes and affine or 3D projection.

Scale Invariant Feature Transform (SIFT) consists of four major stages: scale-space extrema detection, key point localization, orientation assignment and key point descriptor. The first stage identifies key locations in scale space by looking for locations that are maxima minima of a difference-of-Gaussian function [12]. Each point is used to generate a feature vector that describes the local image region sampled relative to its scale-space coordinate frame. Image keys are created from the feature vector that allow for local geometric deformations by representing blurred image gradients in multiple orientation planes and at multiple scales. The keys are used as input to a nearest-neighbour indexing method that identifies candidate object matches. Final verification of each match is achieved by finding a low-residual least-squares solution for the unknown model parameters.

Speeded Up Robust Feature (SURF) is a robust image detector & descriptor, first presented by Herbert Bay [13]. SIFT and SURF algorithms employ slightly different ways of detecting features. SIFT builds an image pyramids, filtering each layer with Gaussians of increasing sigma values and taking the difference. SURF is based on sums of approximated 2D Haar wavelet responses and makes an efficient use of integral images. It calculates determinant of Hessian blob detector, from which it uses an integer approximation. These computations are extremely fast with an integral image.

Random Sample Consensus (RANSAC) proposed by Fischler and Bolles [14] is an iterative method to estimate parameters of a mathematical model from a set of observed data which contains outliers. It is a non-deterministic algorithm which produces a reasonable result only with a certain probability. The probabilities increase as more number of iterations is allowed. In RANSAC a procedure exists which can estimate the parameters of a model that optimally explains or fits in the small data.

SURF is less time consuming than SIFT where as RANSAC is invariant to affine transform. SIFT based methods are expected to perform better for objects with rich texture information as sufficient number of key points can be extracted but require sophisticated indexing and matching algorithms for effective object recognition. An advantage of RANSAC is the ability to do robust estimation of the model parameters, i.e., it can estimate the parameters with a high degree of accuracy even when a significant number of outliers are present in the data set. A disadvantage of RANSAC is that there is no upper bound on the time it takes to compute these parameters and a good initialization is needed.

3. Appearance based approaches: Better discriminating information may reside in the spectral domain or frequency domain. Most recent appearance based techniques involve feature descriptors and pattern recognition algorithms in the frequency domain.

Most widely used approaches perform linear transformations such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). PCA also known as Karhunen-Loeve transformation, most commonly used as dimensionality reduction technique in pattern recognition. It was originally developed by Pearson in 1901 [15] and generalized by Loeve in 1963. PCA does not take class information into consideration. The classes are best separated in the transformed space better handled by LDA, which consider inter cluster as well as intra cluster distances in the classification. Principal Component Analysis is used with two main purposes. First, it reduces the dimensions of data to computationally feasible size. Second, it extracts the most representative features out of the input data so that although the size is reduced, the main features remain, and still be able to represent the original data [15], [16].

A concept of Eigen picture was defined to indicate the Eigen functions of the covariance matrix of a set of face images. Turk and Pentland [17] have developed an automated system using Eigenfaces with the similar concept to classify images in four different categories, which helps to recognize true/false of positive of faces and build new set of image models. For night time detection and classification of vehicle, Thi et al. used Support Vector Machine with Eigenvalue [18]. Sahambi and Khorasani [19] used a neural network appearance based 3D object recognition using Independent component analysis. The Eigenfaces approach has been adopted in recognizing generic objects across different viewpoints and modeling illumination variations [20].

The frequency domain analysis is more attractive as it can give more detailed information about the signal and its component frequencies. Over the past 10 years, the wavelet theory has become one of the emerging and fast-evolving mathematical and signal processing tools for its many distinct merits. Different from the Fast Fourier transform (FFT), the wavelet transform can be used for multi scale analysis of the signal through dilation and translation, so it can extract the time-frequency features of the signals effectively.

Wavelet transforms have been used in the past for time series classification [21]. Originally it was proposed to use DFT to map the time domain function to frequency domain. The wavelet transform [22] is expressed as decomposition of a signal $f(x) \in L^2(R)$ a family of functions, which are translations and dilations of a mother wavelet function $\psi(x)$. The 2D filter coefficients can be expressed as

$$\begin{aligned} h_{LL}(m,n) &= h(m)h(n), \qquad & h_{LH}(k,l) &= h(k)g(l) \\ h_{HL}(m,n) &= g(m)h(n), \qquad & h_{HH}(k,l) &= g(k)g(l) \end{aligned} \tag{2.1}$$

Where, the first and second subscripts denote the low-pass and high-pass filtering respectively along the row and column directions of the image. Wavelet transform can be implemented (convolution and down sample) along the rows and columns separately. 2D discrete Wavelet Transform is performed using low-pass and high-pass filters. After the decomposition four subbands, LL,LH,HL and HH are obtained, which represent the average (A), horizontal (H), vertical (V), and diagonal (D) information respectively. The iteration of the filtering process produces multi level decomposition of an image. Wavelet transforms provide the effective multi scale analysis but are not effective to represent the image with smooth contours in different directions. For acquiring more directional information, Multiscale Geometric Analysis (MGA) tools were proposed such as Curvelet [28], Ridgelet [24], Bandlet [28] and Contourlet [23] etc.

Contourlet transform [23] is a multi scale and directional image representation that uses first a wavelet like structure for edge detection, and then a local directional transform for contour segment detection. A double filter bank structure is used for obtaining sparse expansions for typical images having smooth contours. In the double filter bank structure, Laplacian Pyramid (LP) is used to capture the point discontinuities, followed by a Directional Filter Bank (DFB), which is used to link these point discontinuities into linear structures. The Contourlet have elongated supports at various scales, directions, and aspect ratios. This allows Contourlet to efficiently approximate a smooth contour at multiple resolutions. Nonsubsampled Contourlet was pioneered by Do and Zhou as the latest MGA tool [24] in 2005. Yan et al. [25] proposed a faced recognition approach based on Contourlet transform. Yang et al. [26] proposed a multisensor image fusion method based on Nonsubsampled Contourlet transform. Extensive experimental result show that proposed scheme by Yan's based on Contourlet Transform performs better than the method based on stationary wavelet transforms [23]. Srinivasan Rao, Srinivas Kumar and Chatterji [27] used feature vector using Contourlet Transform for Content Based Image Retrieval System

Candes and Donoho [28] introduced a new multiscale transform named Curvelet transform which was designed to represent edges and other singularities along curves much more efficiently than traditional transforms, i.e., using fewer coefficients for a given accuracy of reconstruction. Implementation of Curvelet transform involves the steps: (1) Subband decomposition, (2) Smooth partitioning (3) Renormalization (4) Ridgelet Analysis. There are two separate Fast Discrete Curvelet Transform (FDCT) algorithms introduced by Starck, Candes and Donoho [29]. The first algorithm is called the Unequally-Spaced Fast Fourier transform (FDCT via USFFT), where the Curvelet coefficients are found by irregularly sampling the Fourier coefficients of an image. The second algorithm is the wrapping transform, which uses a series of translation and a wrap around technique. The wrapping FDCT is more intuitive and has less computation time. Use of the Curvelet Transform for Image Denoising is explained by Starck, Candes and Donoho [29]. A comparative study based on wavelet, Ridgelet and Curvelet based texture classification is well explained by Dettori and Semler [30].

Contourlet transform can represent information better than Wavelet transform for the images having more directional information with smooth contour [23] due to its properties like directionality and anisotropy. Curvelet transform represents edges and other singularities along curves much more efficiently [28]. These two methods have been selected to extract the features for performing the object classification task and also for comparison.

2.1.2 Distance Measures

In order to establish the similarity or closeness of two feature vectors in some feature space, a wide range of distance matrices are used. A distance matrix calculates the distance between two point sets in matrix space [31].

- **Minkowski Norms**

The most commonly used distance matrices are the Minkowski norms. It is defined based on the L_p norm The Norms are popular for their simplicity, speed of calculation and quality of results. Similarity Distance d between two feature vectors is calculated using the following equation:

$$d_P\ (x,y) = ((\sum_{i=1}^{N} |x_i - y_i|^p)^{\frac{1}{p}}) \qquad \textbf{(2.2)}$$

where $x = \{ x_1, x_2, \ldots , x_N)$ and $y = \{ y_1 , y_2, \ldots y_N)$ are the query and targeted feature vectors respectively. N is the number of elements in the vectors.

When p =1, d_1 (x, y) is the city block distance also known as Manhattan distance (L_1)

$$L_1 = d_1\ (x,y) = \left|\sum_{\mathrm{i=1}}^{\mathrm{N}} |x_i - y_i|\right| \qquad \textbf{(2.3)}$$

When p =2 , d_2 (x, y) is the Euclidean distance (L_2) and calculated as

$$L_2 = d_2\ (x,y) = ((\sum_{i=1}^{N} |x_i - y_i|^2)^{\frac{1}{2}}) \qquad \textbf{(2.4)}$$

- **Histogram Intersection**

The histogram intersection is another simple distance matrix that is often used. It was proposed by Swain and Ballard [84]. Their objective was to find known objects within images using color histograms. It is able to handle partial matches when the size of the object with feature vector x is less than the size of the image with the feature vector. The histogram distance d is given as

$$d_{hist}(x,y) = 1 - \frac{\sum_{i=1}^{N} min\ (x_i, y_i)}{\min\ (|x|, |y|)} \tag{2.5}$$

Colors not present in the query image, do not contribute to the intersection distance. This reduces the contribution of background colors. The sum is normalized by the histogram with fewest samples.

- **Bhattacharyya Distance**

A statistical measure known as , Bhattacharyya Distance measure is often used for comparing two probability density functions, which are most commonly used to measure color similarity between two regions [85]. It is very closely related to the Bhattacharyya Coefficient, which is used to measure the relative closeness of the two samples taken into consideration. Bhattacharyya distance can be calculated as

$$d_{bha}(x,y) = \sum_{i=1}^{N} \sqrt{x_i}\ \sqrt{y_i} \tag{2.6}$$

Where x_i and y_i are the probability density function.

- **Cosine Distance**

The cosine distance computes the difference in direction, irrespective of vector lengths. The distance is given by the angle between the two vectors [84]. By the rule of dot product the distance can be calculated using the equation (2.8).

$$x.y = |x|.\ |y| \cos\theta \tag{2.7}$$

$$d_{cos}(x, y) = 1 - \cos\theta = 1 - \frac{x.y}{|x|.|y|} \qquad (2.8)$$

- **Chessboard Distance**

The Chessboard or Chebyshev distance is the maximum distance between the components of two points. This measure creates a space similar to the Manhattan distance but rotated [85].

$$d_{che} = max_i\,(|x_i - y_i|) \qquad (2.9)$$

- **Mahalanobis Distance**

The Mahalanobis distance is a special case of the quadratic-form distance matrices in which the transform matrix is given by the covariance matrix obtained from a training set of feature vectors, that is $A=\Sigma^{-1}$. In order to apply the Mahalanobis distance, the feature vectors are treated as random variables $\mathbf{X}=[\ x_1, x_2, \ldots, x_N]$, where x_i is the random variable of i^{th} dimension of the feature vector [31].

Mahalanobis distance between two feature vector x and y can be calculated as

$$d_{mah} = [(x - y)\Sigma^{-1}(x - y)]^{\frac{1}{2}} \qquad (2.10)$$

In the special case where *xi* are statistically independent, but have unequal variances, Σ is a diagonal matrix as shown in the equation (2.11).

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_N^2 \end{bmatrix} \qquad (2.11)$$

The Mahalnobis distance is reduced to a simpler form [31]:

$$d_{mah} = \sum_{i=1}^{N-1} \frac{(x_i - y_{i)}^{2}}{\sigma_i^2} \tag{2.12}$$

Chi-squared, Kullback-Leibler, Earth Mover's Distance, x^2 Statistics and Quadratic Distances are also well known distance measures used for different applications. The choice of distance metric to be used greatly depends upon application. For general usage, the Minkowski norms will often be a good choice. For applications where speed is preferred over accuracy, the L1 norm or histogram intersection can be used. For applications where the different components cannot be assumed to be independent, a metric such as the Mahalanobis distance may be preferable. In this thesis, Euclidean distance measure has been used for feature matching of training dataset and testing dataset. Bhattacharya distance measure is used for object tracking using the color features.

2.1.3 Performance Matrices

There are many ways of evaluating the performance of a classifier system. A commonly used statistic to measure the performance is 'accuracy'. There are several versions of the accuracy statistics. The basic statistic just measures the percentage of correct classifications out of all the classifications. Accuracy per class and per classification can also be found using statistic [1], [4], [74].

- **Confusion Matrix**

Typical performance measure statistics are calculated using a confusion matrix. This records the true and predicted classification of each object. The two class confusion matrix records four values. The True Positive (TP) value is the number of positive examples correctly classified. Likewise the True Negative (TN) value is the number of negative examples correctly classified. The False Negative (FN) value is the number of positive examples classified as negative and the False Positive (FP) value is the number of negative examples classified as positive.

The User's accuracy (Precision) is the number of correct classifications over all the objects classified as that class.

$$Accuracy = \frac{TP}{TP + FP} \qquad (2.13)$$

The Producers accuracy is also known as recall or sensitivity. Sensitivity is the number of correct classifications over all the objects of that class.

$$\text{Producers accuracy} = \frac{TP}{TP + FN} \qquad (2.14)$$

Specificity measures the proportion of negative examples correctly classified.

$$\text{Specificity} = \frac{TN}{FP + TN} \qquad (2.15)$$

- **Receiver Operating Characteristics Graph**

A Receiver Operating Characteristics (ROC) graph [32] is a visual tool to evaluate classifier performance. A key feature is that it is invariant to class distribution, The ROC graph plots true positive rate against false positive rate. ROC calculates the overall accuracy of a classifier; it does not gauge the accuracy of an individual classification.

- **A Priori and A Posteriori methods**

The work by Giacinto and Roli [33] highlights a priori and a posteriori methods as good confidence estimators. These techniques make use of a validation set. If the k nearest objects in a validation set were correctly classified, then it is likely that the query object also classified correctly. A priori method estimates the confidence without requiring the query to be classified. It simply bases the confidence on how many of the neighbouring objects were correctly classified. A posteriori method requires the query object to be classified first. After that, on the bases of the estimation of confidence on number of the neighbouring objects, the classes are predicted correctly.

2.2 Visual Tracking

Most commonly used visual tracking techniques include Motion Segmentation and Object Tracking. Background subtraction is one of the most common and effective methods of segmenting foreground objects from the background scene. The process of background subtraction involves locating the areas in an image which differ sufficiently, from an image of the background.

A background modeling process has three phases:

1. Model representation – kind of model used to represent the background.

2. Model initialization – Initialization of assumed model.

3. Model adaptation – Mechanism for adapting to illumination changes in the background.

2.2.1 Background model

A number of different approaches to these issues have been proposed. At the naive level, a simple frame difference can be used to obtain the moving objects [34]. In the simplest method, the background model is just the previous frame in the image sequence. The foreground objects are isolated by comparing the difference in color between the current image and the previous image to a threshold value. If the difference exceeds the threshold value, the pixels are referred as part of the foreground. The strength of this approach is the simplicity of the background model. Limited processing power and memory are required to maintain the background model as only the previous frame in the sequence needs to be remembered. It can provide adequate results in situations where the background scene is relatively static but problems arise when the background is constantly changing. In an attempt to address this issue, a number of adaptive background models have been proposed. Considering static motion for the application, simple background approach has been selected.

Background adaptation may be classified as either predictive or non-predictive [35]. Predictive algorithms are known to model the scene as a time series and they would make use of a dynamic model to recover the current input based on past observations. The absolute error between the predicted and the actual observation can then be used as a measure of change. Non-predictive methods try to build probabilistic representation from the observations at a particular pixel. An alternative way for classifying background adaptation methods is either non-recursive or recursive [36].
A non-recursive technique uses a sliding-window approach for background estimation. For non-recursive estimation the L previous video frames are first stored in a buffer and then a background image is constructed making use of the temporal variation of each pixel in the buffer. Non-recursive method requires a large storage memory for slow moving objects. Recursive techniques do not rely on a buffer for estimating the scene. Instead, they recursively update a single or multiple background model based on each input frame. Even though recursive method requires less memory, any error in background model remains for a longer time. To alleviate this problem exponential weighting and positive decision feedback can be used. Non-recursive adaptation techniques include temporal differencing (frame differencing), average filtering, Median filtering and Minimum-Maximum filtering. Recursive techniques on the other hand include Approximated Median filtering, Single Gaussian, Kalman Filtering, Mixture of Gaussians, Clustering based segmentation methods, and Hidden Markov Models.

2.2.2 Statistical Approach

The background as a realization of a random variable with Gaussian distribution (SGM - Single Gaussian Model) is represented by Wren et al.[37] .The mean and covariance of the Gaussian distribution are independently estimated for each pixel in SGM. Stauffer and Grimson [38] presented an adaptive background mixture model for real-time tracking. In their work, they modeled each pixel as a mixture of Gaussians and used an online approximation to update it. The Gaussian distributions of the adaptive mixture models were then evaluated to determine the pixels most likely from a background process, which resulted in a reliable, real-time outdoor tracker which can deal with lighting changes and clutter. A study by Haritaoglu et al. [39] built a statistical model known as w4 which represents each pixel with three values: its minimum and maximum intensity values and the maximum intensity difference between consecutive frames observed during the training period. The model parameters were updated periodically.

Lehigh Omni directional Tracking System (LOTS) presented by Boult et al. [40] is tailored to the detection of non cooperative targets under non stationary environments. This algorithm uses two gray level background images. This allows the algorithm to cope up with intensity variation due to noise or fluttering objects which move in the scene. Each pixel of the input frame is compared to the closest background value and classified as active if the difference exceeds a given threshold.

2.2.3 Object Tracking Techniques

The ability of the model to handle shadows and changing lighting conditions is also increased by utilizing the differing properties of the Color information. Different color spaces also have different advantages when performing background subtraction. The HSV color spaces separate a RGB image into its hue, saturation and value or intensity components. The YC_bC_r color space is widely used for digital video. In this format, luminance information is stored as a single component (Y), and chrominance information is stored as two color-difference components (C_b and C_r). C_b represents the difference between the blue component and a reference value. C_r represents the difference between the red component and a reference value.

Most commonly used visual tracking techniques include Mean Shift Tracking Algorithms [41], [42], Blob Tracking, Particle Filter Algorithms [43], Block Matching [44], [45] and Optical Flow Based Tracking Algorithms [3]. Mean shift tracking algorithm has become popular due to its simplicity and robustness. Tracking is accomplished by iteratively finding the local minima of the distance measure functions using the mean shift algorithm. The mean shift algorithm was originally invented by Fukunaga and Hostetler [46] for data clustering, which they called a "valley-seeking procedure". It was first introduced into the image processing community several years ago by Cheng [47]. Mean Shift based on Color Distribution and Simulated Annealing (SACD-MS), is proposed for human body tracking by Hong [48]. R. Venkatesh Babu [49] proposed a new method to track objects by combining two well-known trackers, Sum-of-Squared Differences (SSD) and color-based mean-shift (MS) tracker. Zoran [50] applied a new 5-Degree Of Freedom (DOF) color histogram based non – rigid object tracking algorithm using Expectation Maximization (EM) Mean shift. Huiyu Zhou [51] proposed the method based on SIFT features and mean shift to track the object. Disadvantage of Mean shift algorithms is to specify the kernel for further tracking. Also some time mean shift tracking algorithms gets stuck at local minima. The similarity measures like Bhattacharya coefficients and Kullback - Leibler divergence are not very discriminative, especially for higher dimensions [52] and difficult to use them due to the sample based calculation for the real time object tracking.

Particle filters [46], [47] are kind of stochastic tracking algorithms that use multiple discrete "particles" to represent the distribution over the location of the target. It has been shown to be very suitable for performing tracking in cluttered environments due to their ability of maintaining multiple hypothesis of probability distribution. More importantly, particle filters exhibit superior characteristic of recovering from the temporary lost track. Sanjeev Arulampalam [53] proposed a method based on Particle Filter using Bayesian Tracking for the Nonlinear/Non-Gaussian tracking problem.

Hybrid tracker [43] using particle filter and Mean Shift are used by Bo Zhang, Weifeng Tian, and Zhihua Jin. Tang Sze Ling [54] described the characteristic of the motion tracker based on color as the key feature to compare the object's similarity for object detection and tracking. Lowe [55] used model based object tracker using Marr–Hildreth edge detector to extract edges from an image. Stanley T. Birchfield and Rangarajan [56] presented a particle filtering framework for region-based tracking using spatiogram. M. A. Zaveri, S. N .Merchant and U. B. Desai [57] proposed a neural-network-based tracking algorithm. Erdem [44] proposed the method based on "defocus energy" which is utilized for automatic segmentation of the object boundary and it is combined with the histogram method to track the object more efficiently. Since particle filters require a large number of particles for accurately representing the probability distribution, it limits their applications to real time occasions.

In the field of Motion Estimation (ME), many techniques have been proposed [58],[59],[60],[61],[62],[63],[64],[65],[66]. Basically ME techniques can be broadly classified as: gradient techniques, pixel-recursive techniques, block matching techniques and frequency-domain techniques.

Among these four groups, block matching is particularly suitable in video compression schemes based on Discrete Cosine Transform (DCT) such as those adopted by the recent standards H.261, H.263 and MPEG family [59],[60]. Block-based motion estimation uses a Block-Matching Algorithm (BMA) to find the best matched block from a reference frame. The basic idea of BMA is to divide the current frame in video sequence into equal-sized small blocks. For each block, we try to find the corresponding block from the search area of previous frame, which "matches" most closely to the current block. Therefore, this "best-matching" block from the previous is chosen as the motion source of the current block. The relative position of these two blocks gives the so-called Motion Vector (MV), which needs to be computed and transmitted. When all motion vectors of the blocks in tracking area have been found, the motion vector happened most frequently is chosen for the correction of tracking area size. Typically, the Sum of Absolute Difference (SAD) is selected to measure how closely two blocks match with each other, because the SAD doesn't require multiplications; in other words, less computation time and resources are needed. There are several methods used to find out the best matching block.

The most commonly used Block Matching Algorithm (BMA) is the Full search (FS)/Exhaustive search (ES), which exhaustively searches for the best matching block within the search window. Full Search Algorithm is the most straight forward strategy. But the computational complexity of Full Search is always too high. As a result Fast Search Algorithm has been developed. In fast BMA using a fixed set of search patterns, the assumption is that, the matching error decreases monotonically as the search moves towards the position of the global minimum error and the error surface is uni-modal. Few fast block matching motion estimation algorithms were Two-Dimensional Logarithmic Search, Three Step Search [61], Four Step Search [62], Block-Based Gradient Descent Search [63], Diamond Search (DS) [64], Cross-Diamond Search (CDS) [65] etc. Adaptive Rood Pattern Search (ARPS) is proposed in [67] to track large motions, with less number of computations by using Zero Motion Prejudgment (ZMP) for the reduction in the computation complexity. Novel Hexagon-based Search (NHS) [66] algorithm has been incorporated in recently developed H.264/AVC video coding standard. These methods are mainly used for image compression but for object tracking. These methods are more time consuming than the standard mean shift and particle filter techniques.

Other object tracking methods involve the shape based and motion based tracking. Cutler and Davis [68] proposed a method that used the periodic shape changes that occur during the walking motion. To analyze the periodic nature of a particular object, its appearance throughout the image sequence must be remembered. A similarity measure between each object image is then generated. If the motion is periodic, this similarity measure will also be periodic as the appearance of objects will repeat. The Fourier transform of the similarity measure can be used to identify peaks in the power spectrum corresponding to the fundamental frequencies of the motion. If a peak exceeds some threshold value, the motion is regarded as periodic. Cutler and Davis also suggest a method to distinguish different types of periodic motion by comparing the similarity images to those generated by a training set. In this fashion, they are able to classify motion as human, animal, or other. While this approach provided reliable classification results, the method is memory intensive, requiring an image of each object in every frame to be stored. Calculating the self similarity between each of these images is also computationally expensive. Another problem with using periodic motion as a classifier is that it is only effective when the subject is moving. If the subject pauses to look at something or talk to another subject, the decision made by the classifier is unreliable.

A method outlined by Lipton [69] uses an optical flow based technique to classify moving objects as rigid or non-rigid. This is achieved using the observation that rigid objects will generate less residual flow than non-rigid objects during non-rotational motion. To calculate the residual flow of a moving body, its net motion, defined as the absolute position change of the object, was determined using a tracking algorithm. The optical flow vector for each pixel in the object was then computed. The residual flow of the body was calculated by subtracting the net motion of the body from the optical flow vector associated with each pixel. Rigid bodies have little residual flow as all pixels that make up the object are moving in the same direction. The optical flow of each pixel is approximately equal to the net motion of the body, resulting in a small residual flow value. Non-rigid objects will display greater residual flow as some pixels that make up the object are moving in different directions to the overall body. The optical flow directions of these pixels are different to the net motion, resulting in a larger residual flow value. Thus, Lipton distinguished between rigid objects such as vehicles and non-rigid objects such as humans.

In the summary with this chapter, background work and literature survey based on the object classification and motion analysis have been described. Based on the literature survey, efficient and computationally fast algorithms are selected for implementation of the proposed algorithm.