Hybrid Gaussian Noise Filtering and Video object Tracking Using Support Vector Machine (SVM) Technique

C. Nithya and P. Vijayakumar

Abstract--- Video Surveillance and their installations are gradually being used to public services and association in order to obtain high level of security. In Video Surveillance applications, real-world Closed Circuit Television (CCTV) footage frequently creates new difficulties to object tracking because of to Pan-Tilt-Zoom operations, low quality of camera and different operational environments. The majority of significant difficulties are moving background, movement blur and rigorous size changes. However in the machine learning based video tracking system users have achieved better performance and exactness of object motion detection when compared to conventional video tracking systems. Particularly Convolutional Neural Networks (CNNs) have attains enhanced performance in object detection and it is being used to follow a more capable object tracking scheme. However CNNs based object tracking scheme becomes very challenging difficult to trace the object for less quality images. To enhance the quality and clarity of CCTV scenes, Motion Adaptive Gaussian Filtering (MAGF) denoising filtering is proposed in this paper, it is applied to three following frames, noise frames and detects the video movement region and still region. MAGF is performed based on the variation among the before and after video frame whether it is on movement or fixed. Within the MAGF, the Temporal Filter will be applied to stationary part of the video frame and Spatial Filter will be applied particularly to the movement part of the video frame in CCTV scenes. By using these filters noises in the video frames are removed and clarity of the video frames is enhanced to CCTV scenes. Then Kernel Support Vector Machine (KSVM) is proposed for multi-object tracking scheme and it is being utilized to follow a more capable object tracking scheme.

In this work, make use of heterogeneous training video frames and video frames augmentation is investigated to enhance their multi-object tracking detection rate in CCTV scenes. Furthermore, it is proposed KSVM to make use of the objects spatial transformation parameters which calculate the development of intrinsic camera parameters and consequently adjust the object detector for higher performance.

Index Terms--- Closed Circuit Television (CCTV), Motion Blur, Pan-Tilt-Zoom (PTZ) Operations, Motion Adaptive Gaussian Filtering (MAGF), Kernel Support Vector Machine (KSVM), Recurrent Convolutional Neural Network (R-CNN), Spatial Transformer, Noise Removal.

I. INTRODUCTION

Video surveillance have been becomes an undergone extraordinary development because of its usage. For instance, the use of CCTV cameras used for public safety in London, United Kingdom (UK) used in the beginning of sixties by means of placing two cameras in Trafalgar Square. In the beginning of nineties, the security communications regarding London grew to a network contains of thousands of cameras. Presently around four million cameras and more than four million cameras are

C. Nithya, Research Scholar, Department of Computer Science, Sri Jayendra Saraswathy Maha Vidyalaya College of Arts and Science, Coimbatore E-mail:mail2nithu92@gmail.com

P. Vijayakumar, Head Department of Computer Application, Sri Jayendra Saraswathy Maha Vidyalaya College of Arts and Science, Coimbatore. E-mail:vijayvigash@gmail.com

used in London, UK according to the report [1]. Similar examples of development have been found in the other countries by means of their reports for Canada [2], China [2], and Australia [3]. Those types of CCTV systems were initiated to support the police and safety personnel in checking crime. The advantage of camera networks is understandable: instead of having safety stationed at every corner, enormous territories be able to be monitored via a few individuals from the control room. Even if an event is not recognized during the time of its event, recorded data have been used to give confirmation of the crime and to recognize executors and losses after the information. The CCTV technology was efficient in the present issue is with the purpose of use of the several cameras are not being monitored by means of safety personnel, and the recorded footage is verified only when an accident or crime event occurs. Even though the cameras are not monitored continuously at live, many of the control rooms have monitor one operator has to multiple camera views during the same time. The facts with the purpose of concentration have been divided among multiple camera views formulate it more possible with the purpose of an interesting event will be losses. Studies related to the Sandia National Laboratories used for the U.S. Department of Energy supported this intuition and demonstrated with the purpose of after only 20 minutes of watching and validating monitor screens, the notice of a large amount individuals drops to fine below suitable levels.

Object tracking have been becomes a most motivated considerable awareness in the research area because of its usage in practical applications and specifically for smart video surveillance solutions. Regardless of the development done in recent years, object tracking methods are not stable sufficient for real-world satisfied from CCTV cameras. In adding together to less quality and changing lighting, occlusions and mixed-up scenes with the purpose of pose tracking difficulties, CCTV footage moreover suffers from movement blur and huge affine transformations appropriate to Pan-Tilt-Zoom (PTZ) operations [4]. Many of the work done in the recent work related to object tracking and object detection methods are paying attention on constructing a robust object appearance model, functioning on handcrafted feature demonstration and classifier creation. On the other hand, many of these classifiers are restricted through their shallow construction at the same time as object appearance differences are difficult and time-varying [5].

Recent development and the usage of deep learning have been also used to object detection and localization schemas with the purpose of perform better than the conventional methods. They depending on automatically learning varied features by the use of multi-layer Convolutional Neural Network (CNN). Each layer consists of varied types of neurons characteristic CNN operations, non-linear filtering and spatial pooling. End-to-end training is second-hand toward automatically study hierarchical and object-specific feature illustrations.

Some literature work done related to the deep learning methods are described as follows: Li et al [6] follows the procedure of CNN to multiple visual object tracking for multiple image cues as inputs. In [7] an ensemble of Deep Neural Network (DNN) learning methods has been integrated with online boosting schema. In [8], a single visual object tracking is proposed via the use of online learning tracker to solve movement blurring problem. Some of the research focuses on the use of auxiliary data to train offline a DNN, and then transmit information to object single object tracking. Fan et al [9] follows the procedure of CNNs with a specific feature extractor in offline training set. In [10] a DNN learning tracking method is introduced with the purpose of make use of stacked denoising autoencoder toward study the general features from a huge number of auxiliary images. In the recent work, Wang et.al [11] introduced the procedure of two-layer CNN to study hierarchical features from auxiliary images, which models complex movement conversion and form deviations. In [12] a deep learning design learns the majority of discriminative features by the use of CNN make use of both the ground truth image samples and the image clarification attained in online manner. On the other hand these methods have some major issues which are described as follows.

Many of the existing object detection methods are focused on creating an object appearance model, working under handcrafted feature illustration and classifier creation. On the other hand, many of these classifiers are restricted because of their shallow structures whereas object appearance distinction is difficult and time-varying. At the same time the quality of images proofed by means of both digital and analogue CCTV systems are differed significantly. Actually, “anecdotal evidence proposes with the purpose of over 80% of the CCTV footage absolute to the police is extreme from perfect; particularly if it is being used for primary identification is being required, for instance, via medium release”.

In this work a new multiple-object detection framework is proposed for tracking by the use of detection applications with the purpose of deal with the challenges of real-world CCTV videos.

To increase the quality of CCTV scenes, Motion Adaptive Gaussian Filtering (MAGF) analyzes is proposed comparing three successive frames, noise frames and identifies the motion area and still area. Better detail will remain on the stationary details and less ghosting on the moving object in the image. Moreover, graph partitioning based video segmentation algorithm partitions an input video into several frames. Each frame is heterogeneous with respect to one or more properties i.e. the variation of measurements within the regions should be considerably less than variation at the object borders. Kernel Support Vector Machine (KSVM), which offers state-of-the-art performance in object detection, is increasingly utilized to pursue a more efficient tracking scheme.

To enhance the quality and clarity of CCTV scenes, Motion Adaptive Gaussian Filtering (MAGF) denoising filtering is proposed in this paper, it is applied to three following frames, noise frames and detects the video movement region and still region. By using MAGF noises in the video frames are removed and clarity of the video frames is enhanced to CCTV scenes. Furthermore, graph partitioning based video segmentation is also proposed this work that partitions the input video into several frames. Each frame is heterogeneous that relates to one or more properties *i.e.* the difference of measurements inside the regions must be significantly less than difference at the object boundaries. Then Kernel Support Vector Machine (KSVM) is proposed for multi-object tracking scheme and it is being utilized to follow a more capable object tracking scheme. In this work, the make use of heterogeneous training video frames and video frames augmentation is investigated to enhance their multi-object tracking detection rate in CCTV scenes.

II. LITERATURE REVIEW

This section provides a detail review of object detection methods and classification methods related to object detection and tracking in common, and then focal point on humans in together single frame and in multiple video frame. They also discuss the some major characteristics of the different algorithms, applicability, and its issues.

A. *Review Object Detection Methods in CCTV*

In general object detection in a video and frames deals with problem of detecting and locating examples of related objects in a video scene via matching video features established in the image to object features, and is perform a classification task to those detected objects. In common there might be more than one data object in the scene, and these objects might be present anywhere in the scene, so new methods is required for object detection. Another important step is object recognition thus finding an exacting object by means of differentiating between groups of data objects in the similar class [13] by finding its pose. The major different among object detection and object recognition is described as follows. However the object detection is performed based on the inference on image features, whereas the object recognition furthermore considers the higher level object concepts and analysis. However both object detection and object recognition tasks might uses a some object features those are movement, texture, colour and shape. Normally in an object recognition task there is a database of objects beginning which you would have to discover the closest match toward the present object.

Among these two task object detection task should be used in several application by considering a single video or image as input for image database retrieval, or image series as in video for automatic target detection and object tracking in visual surveillance and video surveillance applications. In the recent work survey studies the information of three major classification schemes for object detection and tracking in images. The recent object detection schemas were performed based on the computer vision technologies and experimented to single image snapshot, conversely there are few more techniques also used from pattern detection and statistical signal processing. The survey describes the details of three major object detection methods namely, feature-based methods, Movement-based recognition [29] and model-based recognition [14]. Movement based recognition techniques make use of intrinsic movement properties of the object for object detection, for instance the step of a walking person.

Model-based recognition technique makes use of two dimensional and three dimensional models for object detection, simultaneously combining the procedure of movement model and poses constraints. For example consider a VIEWS system [15] at the University of Reading is a three-dimensional schema for vehicle object tracking. The Finder system [16] is used to improve three-dimensional description of a person in big room. It tracks a single non occluded person in difficult scenes in a video, and has been also applied to several applications. However the feature-based object recognition is performed based on the single snapshot images that finds a different feature set moreover in the image space or in an appropriate feature space.

Some of the feature-space methods consist of wavelets domain, eigen space [17], multi-dimensional histogram feature dimension space [18], and shape space. These methods different features they are intensity, directional intensity gradients, colour, texture and wavelet coefficients are used towards explain the object in image space. In feature-space methods, two most important methods are vision based methods and pattern recognition based methods is normally used for object detection. Vision based methods needs examination and extraction of object features, and object detection is obtained by differentiation of object from other classes. In pattern recognition methods follows the procedure of some criteria’s are minimum variance and minimum number of inequitable components might be used to extract features which are then passed to a classification algorithm that extract structural information.

Some of the other object detection methods are normal classifiers, and patch based classifiers. Normal classifiers plan on creating statistical association among objects and its features. Patch based classifiers are performed based on the hand detects objects via investigative a patch of a frame designed for verification of the object. These statistical classifiers are categorized into three types are generative, registration, and discriminative approaches.

The generative approach aims to distinguish particularly helpful object features and their spatial associations [20], and then recombine those features in an identified manner toward create a new object model. Examples include Bayesian Networks [21], and cluster-based models. The registration approach aims to support and match related feature points among two or more images [22] in stereo imaging technology that results in difference maps from which objects are detected. The discriminative approaches aim to classify objects by means of generic descriptors via learning a discriminating function. Patch classifier model initially extract some image features from video or image and construct a classification model to these features. The resulting classifier must describe various individually identifiable set of features from the fundamental patch.

Patch classifiers follows a some conventional classifiers such as Support Vector Machine (SVM), Adaboost, and Feed Forward Neural Networks (FFNNs) [22]. These machine learning methods are used to study the fundamental structure of an object. These classifiers which generalize by the use of learning object features in order to differentiate the objects from other classes. ANN is self organizing structures are capable to fine-tune itself subsequent to getting inputs from its location. It is a non linear network designed for approximating functions toward some random level of accuracy.

B. Review of Multiple Objects Tracking

Object tracking involves concerning the related object in succeeding frames ultimately. Tracking then communicate objects transversely frames. In the later case an object and its association is together predictable by iteratively revising object position and calculate object features among successive frames. In general object tracking algorithms have been categorized into single object tracking and multiple object tracking. In single object tracking algorithms, the interactions among single object and background is measured to video scene with high complexity. In multiple objects tracking algorithms, added interactions among objects should be considered. This makes multiple object tracking algorithms become more difficult and challenging particularly in connecting measurements to model predictions. In the recent work several numbers of multiple objects tracking algorithms is proposed to find target tracking area.

Based on the literature and the recent work multiple object tracking algorithms have been categorized into feature-based, model based, region based and contour based tracking algorithms [23]. Another classification related to [24] is by structure of feature representation or how feature association problem is solved. Under structure of feature representation three general types such as point tracking, kernel tracking, and silhouette tracking are followed in the recent work.

Point tracking [25] is the association of detected objects denoted as point features across frames. Point trackers are appropriate for tracking objects of each and every one size. Regularly multiple points are required to track extremely large objects. Kernel tracking related to association of objects across frames by means of using rectangular, elliptical templates. However the major issue of multiple object tracking is that the data connection, how to obtain optimal mapping among actual measurements and predicted measurements. Thus these data associations problems are solved by using Joint Probabilistic Data Association Filter (JPDAF) [26], Monte Carlo data association filter [27], and nearest neighbour filter schemas. The optimum data association follows the procedure of several filters thus provides the results for the creation of track introduction, track execution, track maintenance, precise modeling of false measurements, and modeling of individuality measurements.

III. PROPOSED METHODOLOGY

Machine learning video tracking system, users have achieved better performance and exactness of object motion detection when compared to conventional video tracking systems. Convolutional Neural Networks (CNNs) have attains enhanced performance in object detection and it is being used to follow a more capable object tracking scheme. However CNNs based object tracking scheme becomes very challenging difficult to trace the object for less quality images. To enhance the quality and clarity of CCTV scenes, Motion Adaptive Gaussian Filtering (MAGF) denoising filtering is proposed in this paper, it is applied to three following frames, noise frames and detects the video movement region and still region. Furthermore, graph partitioning based video segmentation is also proposed this work that partitions the input video into several frames. Then Kernel Support Vector Machine (KSVM) is proposed for multi-object tracking scheme and it is being utilized to follow a more capable object tracking scheme. Furthermore, it is proposed KSVM to make use of the objects' spatial transformation parameters which calculate the development of intrinsic camera parameters and consequently adjust the object detector for higher performance. Figure 1 illustrates the architecture diagram representation of entire work.

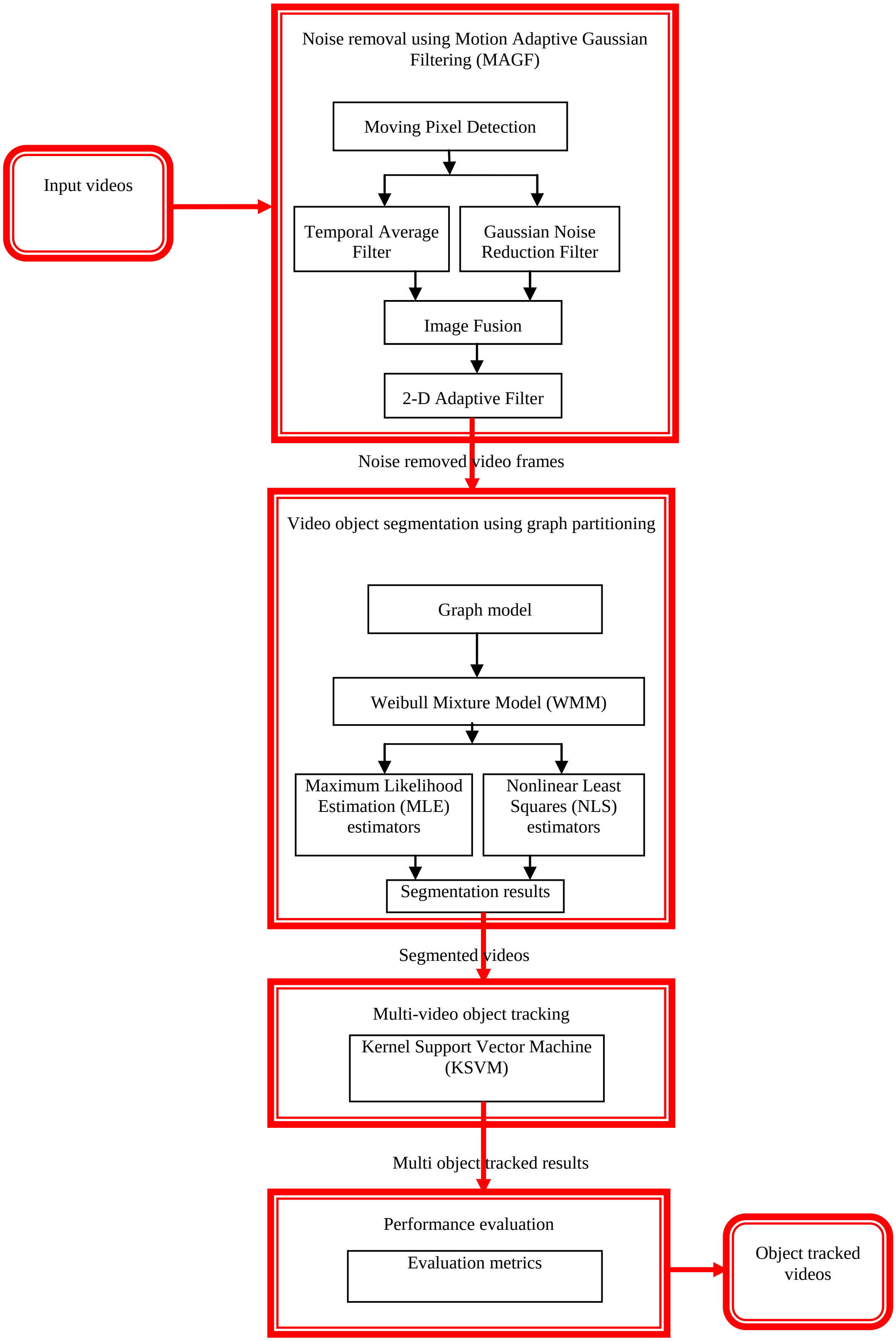
Input videos
Noise removal using Motion Adaptive Gaussian Filtering (MAGF)
Moving Pixel Detection
Temporal Average Filter
Gaussian Noise Reduction Filter
Image Fusion
2-D Adaptive Filter
Noise removed video frames
Video object segmentation using graph partitioning
Graph model
Weibull Mixture Model (WMM)
Maximum Likelihood Estimation (MLE) estimators
Nonlinear Least Squares (NLS) estimators
Segmentation results
Segmented videos
Multi-video object tracking
Kernel Support Vector Machine (KSVM)
Multi object tracked results
Performance evaluation
Evaluation metrics
Object tracked videos

Figure 1: Architecture Diagram

A. Noise Removal Using Motion Adaptive Gaussian Filtering (MAGF)

In this work, a new Motion Adaptive Gaussian Filtering (MAGF) is proposed for noise removal of CCTV footage images that achieves enhanced performance with high quality and clarity of CCTV footage with less computation requirement. CCTV footage images are used concurrently to divide the static regions and dynamic regions, where different noise reduction filters are used by MAGF. Within the MAGF, there are two types of filters namely the Temporal Filter will be applied to stationary part of the video frame and Spatial Filter will be applied particularly to the movement part of the video frame in CCTV scenes. For CCTV footage images, make use of spatio-temporal noise reduction filter [28] that efficiently remove noise in the video frames by means of considering movement variation. For this purpose define two videos, including an input CCTV footage frames $g(x, y, t)$ and a noise removed video$g(x, y, t + 1)$, are considered at the same time. Five major important steps are carried out in the proposed MAGF are: Moving Pixel Detection, Temporal Average Filter, Gaussian Noise Reduction Filter, Image Fusion, and 2-D Adaptive Filter in Static Regions.

Moving Pixel Detection initially divides the video frames into static regions and dynamic regions. Next, the noises presented in the static regions should be reduced by using temporal Average Filter, and the Gaussian Noise Reduction Filter with movement part is applied to the dynamic regions for noise removal. To combine both static region and dynamic region noise removed frames image fusion method is used in this work.

At finally two-dimensional adaptive filter is applied to static regions that further decrease the noise in the static regions. The details of working procedure of each operation are described as follows. At initially describe the Mean Absolute Difference (MAD) of a correlation window approximately a pixel (x, y) with movement vector (u, v) as

$$MAD_{(x,y,t)}(u,v) = \frac{1}{(2M+1)(2N+1)} \sum_{m=-M}^{M} \sum_{n=-N}^{N} |g(x+m, y+n, t) - g(x+m+u, y+n+v, t+1)| \tag{1}$$

Where $(2M + 1)$ and $(2N + 1)$ are the width and height of the correlation window. Through $MAD_{(x,y,t)}(0, 0)$, divide dynamic regions and static regions at the same time record the locations of dynamic pixels in the “moving pixel mask,” $mp(x, y, t)$:

$$mp(x,y,t) = \begin{cases} 0 \ \ if \ MAD_{(x,y,t)}(0,0) < mth \\ \quad 1 \ else \end{cases} \tag{2}$$

Where *mth* is a threshold value and it is directly proportional to the noise standard deviation. The experimentation value of threshold (*mth)* is equals to $1.5\sigma_\eta$. An example of the moving pixel mask is carryout by using VOC2007 video sequence. Note that, the noise variance $\sigma^2{}_\eta$ have been determined by using noise variance estimation [28], or it has been also computed iteratively from the frame variation of the static regions. A simple temporal average filter have been applied in this paper work to reduce the noises in the video frames i.e., for $mp(x, y, t) = 0$,

$$ta(x,y,t) = \frac{1}{2}[g(x,y,t) + g(w,y,t+1)] \tag{3}$$

By means of which the noise variance is decreased to $\sigma^2{}_{\eta/2}$. Conversely, in the dynamic regions, the dense movement field is first determined using the following equation,

$$\left(U_{(x,y,t)}, V_{(x,y,t)}\right) = \underset{(u,v)}{\arg\min} \, MAD_{(x,y,t)}(u,v) \tag{4}$$

Where $(U_{(x,y,t)}, V_{(x,y,t)})$ is the movement frame vector of the pixel (x, y). Next, a spatio-temporal filter has been used with movement compensation in Gaussian Temporal Noise

Reduction Filter step. Here, a three dimensional adaptive filter is extended from two dimensional adaptive filter is described as follows.

$$\mathrm{gt}(x,y,t) = g(x,y,t) - \frac{\min(\sigma_{tl}^2, \sigma_n^2)}{\sigma_{tl}^2}[g(x,y,t) - \mathrm{mt}(x,y,t)] \tag{5}$$

If the video frames contains any temporal noise it have be represented as independent additive Gaussian noise using the following equation:

$$g(x,y,t) = f(x,y,t) + n(x,y,t) \tag{6}$$

where f(x, y, t) is the original video frames pixel intensity at location (x, y) at time t, n(x, y, t) is the Gaussian noise with zero mean and σ^2_η variance, and g(x, y, t) is the noise video frames pixel intensity.

$$\mathrm{mt}(x,y,t) = \frac{1}{2(2R+1)(2S+1)}\left\{\sum_{m=-R}^{R}\sum_{n=-S}^{S}\left|g(x+m,y+n,t) + \sum_{m=-R}^{R}\sum_{n=-S}^{S}|g(x+m+U,y+n+V,t+1)|\right|\right\} \tag{7}$$

$$\begin{aligned}\sigma_{tl}^2 = {} & \frac{1}{2(2R+1)(2S+1)} \\ & \times \{\sum_{m=-R}^{R}\sum_{n=-S}^{S}\left|[g(x+m,y+n,t) - \mathrm{mt}(x,y,t)]^2\right. \\ & \left. + \sum_{m=-R}^{R}\sum_{n=-S}^{S}|g(x+m+U,y+n+V,t+1) - \mathrm{mt}(x,y,t)^2|\right|\}\end{aligned} \tag{8}$$

Note that, only video frames pixels with mp(x, y, t) = 1 is considered for together motion estimation and Gaussian Filtering (GF) with video frames, it saves computation time of preprocessing methods. In Image Fusion, the two images, $\mathrm{ta}(x,y,t)$ and $\mathrm{gt}(x,y,t)$, are then fused together to form noise removed image f'(x,y,t) using the following equation.

$$f'(x,y,t) = \begin{cases} ta(x,y,t)\ if\ mp(x,y,t) = 0 \\ gt(x,y,t) else \end{cases} \tag{9}$$

Following that, in two dimensional Adaptive Filter in Static Regions step, because the noise variance of the static regions in $\hat{f}(x,y,t)$ is lesser than with the purpose of g(x, y, t), 2-D adaptive filter might perform well, as shown in the following equations.

$$\hat{f}(x,y,t) = \begin{cases} f'(x,y,t)\ if\ mp(x,y,t) = 1 \\ f'(x,y,t) - \dfrac{min\left(\sigma_L^2, \frac{\sigma_n^2}{2}\right)}{\sigma_L^2}\ if\ mp(x,y,t) = 1 \\ [f'(x,y,t) - m(x,y,t)]\ else \end{cases} \tag{10}$$

Where

$$m(x,y,t) = \frac{1}{(2R+1)(2S+1)} \times \sum_{m=-R}^{R}\sum_{N=-S}^{S} f'(x+m,y+n,t) \tag{11}$$

$$\sigma_L^2 = \frac{1}{(2R+1)(2S+1)} \times \sum_{m=-R}^{R}\sum_{N=-S}^{S}[f'(x+m,y+n,t) - m(w,y,t)]^2 \tag{12}$$

The final output restored image is $\hat{f}(x,y,t)$.

B. *Video Segmentation Using Graph Partitioning*

The improved significance of video examination requires well-organized video segmentation, because of high processing cost of video data. Video segmentation provides a better solution which effort to cluster related pixels simultaneously under a spatio-temporal setting, moreover by methods with the purpose of create a set of hierarchical segmentations [29], or by methods with the purpose of group superpixels simultaneously toward form spatio-temporal superpixels [30].

In this work we proposed a new grouping based video segmentation method that increase the results of video object tracking when compared to state-of-the-art by as much as 30%, and is roughly 20 times faster. Hierarchical video segmentation methods have provides better results on normal datasets such as such as Segtrack [31] and Chen's Xiph.org [32] using a variety of metrics . On the other hand, their applicability in video processing applications still becomes restricted, because of high computational complexity and the automatically selection of suitable hierarchical layer for specific applications have been becomes also difficult task. To solve these issues, the recently Uniform Entropy Slice (UES) method ([33]) is proposed that chooses the diverse supervoxels from the many hierarchical layers into form single output segmentation through balancing the amount of video frames information of the chosen supervoxels [31]. Furthermore, in this research work a new graph partitioning based video segmentation algorithm is proposed that divides the input video files into several frames.

Each frame in the video is heterogeneous with respect to more properties *i.e.* the difference of measurements inside the regions must be significantly less than variation at the object borders. Let us consider $\mathrm{G} = (\mathrm{V}, \mathrm{E})$ graph model with an edge e $u, v \in E$ connects two neighboring nodes, $\mathrm{v} \in \mathrm{V}$. Let x_i is considered as the weight of the i^{th} edge e_i in the graph G, the task is towards allocate a binary class label to e_i by means of an objective function $y_i = I(x_i)$, such with the purpose of $y_i = 1$ if e_i is an intra-cluster edge with the purpose of must be retained, or $y_i = 0$ if e_i is a inter-cluster edge with the purpose of must be removed from the graph model G. For a given video frame feature $f \in F$, x_i is the similarity distance among the video frame feature histograms of nodes v_a and v_b associated by means of the i^{th} edge e_i such that $\mathrm{x_i} = \mathrm{D}(\mathrm{v_a}, \mathrm{v_b} | \mathrm{f})$, and represent x as similarity distances of diverse features.

Neighboring nodes in a temporal graph are defined as nodes with the purpose of they are spatially or temporally neighboring to each other, where temporal adjacency in this graph partitioning framework is defined another way relying on whether the movement feature is used: if the movement feature is used, the temporal neighbors of v_a are nodes positioned inside a $\mathrm{n} \times \mathrm{n}$ window on the next temporal frame, where the center of the window is specific by means of the mean movement vector of v_a; if the movement feature is not used, then temporal adjacency is described by a $4\mathrm{n} \times 4\mathrm{n}$ window straightforwardly on the subsequently temporal frame by using the centroid of v_a as the center of the window. Because the edges have simply be intra-cluster or inter-cluster, the distribution of the edge weights x determined from a given video frame feature f is consequently collected of the two individual populations, where the lower distance values of video frames features more related to be intra-cluster distances and the higher distance values are more related to inter-cluster group. But in general k means clustering methods have some major disadvantages. So new Lp-norm based distance (e.g. Earth Mover's Distance [35]) measure is introduced in this work for the measuring the similarity between video frame feature histograms, with the purpose of Lp distance follows the procedure of Weibull distribution, if the distance between the two video frame feature vectors to be compared are correlated and non-identically circulated. It is consequently hypothetically possible to discover the critical value by means of appropriate a 2-component Weibull Mixture Model (WMM) for L_p distance statistics, and maintain the cross-point of the two components as the critical value for graph partitioning. The WMM is generally described as follows:

$$W^{k}(x|\theta) = \sum_{k=1}^{K} \pi_k \varphi_k\,(x, \theta_k) \tag{13}$$

$$\varphi(x|\alpha, \beta, c) = \frac{\beta}{\alpha}\left(\frac{x-c}{\alpha}\right)^{\beta-1} e^{-\left(\frac{x-c}{\alpha}\right)^{\beta}} \tag{14}$$

where $\theta_k = (\alpha_k, \beta_k, c_k)$ is the parameter vector for the k^th mixture component, and φ denotes the three-parameter Weibull Probability Density Function (PDF) among the scale (α), shape (β), location (c) parameter, and mixing parameter π such with the purpose of $\sum_k \pi_k = 1$. In this case, the two-component WMM contains a 6-parameter vector $\theta = (\alpha_1, \beta_1, \alpha_2, \beta_2, c_2, \pi)$ with the purpose of yields the following complete form:

$$W^{2}(x|\theta) = \pi\left(\frac{\beta_1}{\alpha_1}\left(\frac{x}{\alpha_1}\right)^{\beta_1-1}\right) e^{-\left(\frac{x}{\alpha_1}\right)^{\beta_1}} + (1-\pi)\left(\frac{\beta_2}{\alpha_2}\left(\frac{x-c_2}{\alpha_2}\right)^{\beta_2-1}\right) e^{-\left(\frac{x-c_2}{\alpha_2}\right)^{\beta_2}} \tag{15}$$

In order to optimize the above mentioned WMM, need to define the parameters of WMM by using Maximum Likelihood Estimation (MLE) and Nonlinear Least Squares (NLS) estimators. The log-likelihood function of $W^{2}(x|\theta)$is given by:

$$\ln \mathcal{L}(\theta: x) = \sum_{n=1}^{N} \ln\left\{\pi\left(\frac{\beta_1}{\alpha_1}\left(\frac{x}{\alpha_1}\right)^{\beta_1-1}\right) e^{-\left(\frac{x}{\alpha_1}\right)^{\beta_1}} + (1-\pi)\left(\frac{\beta_2}{\alpha_2}\left(\frac{x-c_2}{\alpha_2}\right)^{\beta_2-1}\right) e^{-\left(\frac{x-c_2}{\alpha_2}\right)^{\beta_2}}\right\} \tag{16}$$

Nonlinear Least Squares (NLS) estimated x with histograms where the suitable bin-width is adaptively computed by $l = 2(IQR)^{n-1/3,}$ where IQR is the interquartile range of x via sample size n [35]. Subsequently, NLS optimizes the parameter value of ELM in equation (16) by considering the height of each bin as a curve fitting problem; the least squares minimize is determined via the use of proposing a new trust-region method [36]. Both of the above mentioned estimators require an initial guess parameter vector θ' to solve local optimal problem. Known a node v_i and its group of adjacent neighbors vN_i, the neighbor with the purpose of is mainly related to v_i is most related to the same cluster as v_i. So the use of Weibull Probability Density Function (PDF) over the minimum neighbor distance of all superpixels becomes very useful for video segmentation. For the first guess of the second mixture component, mine the edge weights with the purpose of there are more than a number of percentile of x, where p = 0.6 was found to be a good point in this experimentation work.

C. Proposed Kernel Support Vector Machine (KSVM)

SVM [37] is learning machines that plot the training video frames in high dimensional video feature space, labeling each frame vectors by its class. SVMs classify video frame by determining a set of support vectors, which are members of the set of training video frame inputs that outline a hyper plane in the video feature space. SVMs provide a generic mechanism to fit the surface of the hyper plane to the video frames through the use of a kernel function. The number of free parameters used in the SVMs depends on the margin that separates the video frames but not on the number of input features. The preliminary step is to label the video frames in noise removed CCTV footage file. Two class labels are used namely 1 for multi-object tracking and -1 for no multi-object tracking. This labeled file is used for training the SVM. A set of CCTV footage is taken as testing video frames. The training video frames consist of all the possibilities of multi-object tracking and no multi-object tracking class types so that multi-object tracking can be done. The output is written in two files one containing only multi-object tracking class and the other contains no multi-object tracking class. To keep the computational load reasonable, the mappings used by SVM schemes are designed to ensure that dot products may be computed easily in terms of the variables in the original space, by defining them in terms of a kernel function $\mathrm{K(x,y)}$ selected to suit the problem. Given some CCTV footage image samples I, a set of n frames of the form

$$I = \{x_i, y_i\} | x_i \in \mathbb{R}^p, y_i \in \{-1,1\}_{i=1}^{n} \tag{17}$$

where the $\mathrm{y_i}$ is either multi-object tracking (1) or no multi-object tracking class(-1), indicating the class to which the point $\mathrm{x_i}$ belongs. Each $\mathrm{x_i}$is a p-dimensional video frames vector. Find the maximum-margin hyperplane that divides the video frames having $\mathrm{y_i} = 1$ from those having $\mathrm{y_i} = -1$. Any hyperplane can be written as the set of video frames x satisfying. Maximum-margin hyperplane and margins for an SVM trained with video frames from two classes. Video frames on the margin are called the support vectors.

$$w.x - b = 0 \tag{18}$$

Where denotes the dot product and w the normal vector to the hyperplane. The parameter $\frac{\mathrm{b}}{||\mathrm{w}||}$ determines the offset of the hyperplane from the origin along the normal vector w. If the training video frames are linearly separable be able to choose two hyperplanes in a way that they divide the video and there are no video frames between them, and then try to maximize their distance. The region bounded by them is called "the margin". As prevent video frames from falling into the margin by considering following constraint: for each i either $\mathrm{w.x_i - b \geq 1}$ of the multi-object tracking class 1 or $\mathrm{w.x_i - b \leq -1}$ for the no multi-object tracking class. It is defined as $\mathrm{y_i(w.x_i - b) \geq 1}$ for all $1 \leq \mathrm{i} \leq \mathrm{n}$, Minimize $\mathrm{(w,b)}$ $\left\|\mathrm{w}\right\|$ subject to (for any $\mathrm{i} = 1,..\mathrm{n}$) $\mathrm{y_i(w.x_i - b\,) \geq 1}$. If the kernel used is a Gaussian radial basis function, the corresponding video frame feature space is a Hilbert space of infinite dimensions. Maximum margin classifiers are fine regularized, consequently the unbounded dimensions do not spoil the results.

$$k(x,y) = \left(\sum_{i=1}^{n} x_i y_i + c\right)^2 = \sum_{i=1}^{n} (x_i^2)(y_i^2) + \sum_{i=2}^{n}\sum_{j=1}^{i-1} (\sqrt{2x_i}\, x_j)\,(\sqrt{2y_i}\, y_j) + \sum_{i=1}^{n} (\sqrt{2c}\, x_i)(\sqrt{2c} y_i) c^2 \tag{19}$$

The original concept of SVM is proposed for binary classification [38]. Given a training video frames $(x_i, y_i), i = 1, 2, \ldots, n, x_i \in R^d$, where $\mathrm{x_i}$ is the ith input video feature vector of d-dimension, $y_i \in \{-1, +1\}$ is the corresponding class label, that is 1 for multi-object tracking and -1 for no multi-object tracking , n is the number of training video frames, SVM constructs a separating hyperplane that separates the training video frames vectors perfectly with the closest training video frames vectors beside the hyperplane as far as possible away from those in the other class. This amounts to solving the following optimization problem

$$\min_{\omega,b} \frac{1}{2} \left\|\omega\right\|^2 \; s.t\; y_i(\langle \omega, x_i \rangle + b\,) \geq 1 \tag{20}$$

where $\boldsymbol{\omega}$ and b are the predefined parameters in the hyperplane $< \omega, x + b >= 0$. In the real world applications, conversely, most problems are nonlinear [39]. In this case, the nonlinear video frames should to be mapped to a new kernel space and allow for wrongly tracked multi-object video frames. Using the Lagrange method, the optimization problem (20) in the nonlinear case with a hyperplane margin have been transformed into the dual form and it is described as follows,

$$\min_{\alpha} \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j\, y_i y_j \langle \Phi(x_i), \Phi(x_j) \rangle \;\; s.t. 0 \leq \alpha_i \leq C, \sum_i \alpha_i y_i = 0 \tag{21}$$

Where α_i is the Lagrange multiplier, $\Phi(\boldsymbol{x})$ is the nonlinear mapping function and C is the penalty coefficient. The final decision function is described as follows,

$$Class(x) = sign(\sum_i \alpha_i y_i < \Phi(x_i), \Phi(x) > \; +b) \tag{22}$$

The inner product in the new video frame feature space $\mathrm{K(x, y)} =< \Phi(\mathrm{x}), \Phi(\mathrm{y}) >$ is called kernel function. Moreover the linear kernel which corresponding video frames feature space is just the original training video frames, Radial Basis Function (RBF) kernel is the most commonly used kernel function [39]. Its expression is described as follows,

$$K(x, y) = e^{-\frac{||x-y||^2}{2\sigma^2}}, \sigma > 0 \tag{23}$$

KSVM is first proposed that selects frames in the way of backward elimination. Particularly, in each iteration with the purpose of selects frames which influences the least value of the objective function.

The objective function is J = ½||ω||^2 according to (21). According to Optimal Brain Damage (OBD) algorithm, the change of the objective function with respect to the removing of the ith frames need to satisfies the following expression,

$$\Delta J(i) = \frac{\partial J}{\partial \omega_i} \Delta\omega_i + \frac{\partial^2 J}{\partial \omega_i^2} (\Delta\omega_i)^2 \tag{24}$$

$$\Delta J(i) = (\Delta\omega_i)^2 \tag{25}$$

can disregard the first order term of (23) at the optimum of J, which leads to high object tracking and detection. Therefore remove noise frames iteratively in terms of the absolute or squared value of ω_i as $\Delta\omega_i$equals ω_i in the case of removing the i^{th} frames.

KSVM Classifier to Multi-object target Tracking and Detection

```
Input: Number of the video frames VF for KSVM classifier
Output: Classification result and detection of objects
Procedure KSVM (VF) // input training video frames VF results for the KSVM classifier to multi-object target tracking and detection
Begin
Begin
Initialize C=0 //initially the class labels should be zero
Get input file with video frames VF for training //.
Read the number of input video frames VF from original CCTV footage
$(\mathrm{x_i}.\mathrm{w} + \mathrm{b})\mathrm{k_i} = 0$ // input video frames VF is represented as matrix and denoted by $\mathrm{x_i}$ and w is the weight value matrix whose product is summed with b bias value to give the class value.
$(\mathrm{x_i}.\mathrm{w} + \mathrm{b})\mathrm{k_i} = 1$ // This above equation marks a central classifier margin. This can be bounded by soft margin at one side using the following equation.
Decision function $\mathrm{f(W)} = (\mathrm{x_i}.\mathrm{w} - \mathrm{b})\mathrm{k_i}$ //decision function f(w) decides the class labels for the SVM classification training videos ,
If $\mathrm{f(W)} \geq 1$ for $\mathrm{x_i}$ is the first class // if the F(w) is greater than or equal to the 1 is labeled as first class (tracked videos )
Else
$\mathrm{f(W)} \leq -1$ for $\mathrm{x_i}$ is the first class // if the f(w) is less than or equal to the value of -1 is labeled as second class
The prediction result for (i=1,…n) //after the classification result are performed then check the classification result by testing phase it is check the below function
$\mathrm{y_i}(\mathrm{x_i}.\mathrm{w} - \mathrm{b}) \geq 1$ //if the function is greater than one the results or classified video as predicted (non accepted video)
Display the result //finally we display the classification result
```

IV. Simulation Results

In this section, the simulation setup for the evaluation of the above multi-video object tracking methods such as NNs, CNNs and KSVM is being illustrated. Specified with the purpose of pedestrians is the most important object class of interest, the simulation results will focus on the pedestrian detection without loss of their generality. For this reason a VOC2007 have been selected and applied to multi-video object tracking methods such as NNs, CNNs and KSVM in training phase. VOC2007 [40] is used as a standard dataset for video object tracking with 20 number of classes together with the class person. The ETH dataset [41] is moreover used toward expands the fine-tuning dataset. It consists of annotated pedestrians on a public road. Lastly, a group of videos from the Metropolitan Police of London (MET) 2011 is also used for qualitative experimentation.

A. *Dataset Description*

The initial set of experimentation refers to the investigation of training video frames extension strategies. However the multi-video object tracking methods such as NNs, CNNs and KSVM is fine-tuned by using different training and testing sets. Training through VOC2007 (-10000 object instances) is labeled as VOC and it is used as a baseline for performance. The training set is suffused through series from the ETH dataset, namely “Bahnhof” series (-7500 object instances) labeled as [BAH] and “Sunny Day” (-1900 object instances), labeled as [SUN].

Generally the datasets are divided into two phases namely training and testing set with equal size. 50% dataset samples are used for training set in validation purposes. The assessment of the trained classifiers is performed based on individual testing sets with the purpose of consists of an ensemble of the VOC and ETH testing sets, correspondingly. Training by means of the VOC2007 dataset is used as baseline. The dataset is then enlarged via blurred instances of the VOC2007 dataset, creating [VOC5] with l = 5 pixels and [VOC10] with l = [5, 10] pixels motion blur. The proposed classifiers and existing classifiers are tested on each and every one testing sets, named NoBlur, Blur5px and Blur10px correspondingly. From the results it concludes that the proposed KSVM classifier have achieved higher performance, since the blurred frames is enhanced by using filtering. Conversely, expanding the training video frames with blurred video frames is building the detector stronger, even on non-blurred video frames. Figure 1 show the CCTV video input image with the evaluation of the trained classifier models are performed based on separating testing sets with the purpose of consists of ensemble of the VOC and ETH testing sets, correspondingly. In this work, make use of heterogeneous training video frames and video frames augmentation is investigated to enhance their multi-object tracking detection rate in challenging CCTV scenes. Furthermore, it is proposed KSVM to make use of the objects’ spatial transformation parameters which calculate the development of intrinsic camera parameters and consequently adjust the object detector for higher performance. The KSVM based multiple object visual tracking for CCTV footage performs better than the conventional object tracking method, since the proposed KSVM work is performed based on features and noises in the CCTV footage of the object is removed using filtering with different noisy environments. Figure 2 shows the input CCTV footage representation.

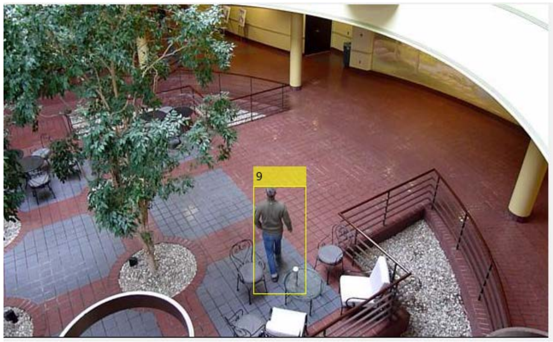

Figure 2: Input CCTV Video Image

Consequently, the effect of expanding the video frame with motion blur is validated and the predicted results are illustrated in the Figure 3.

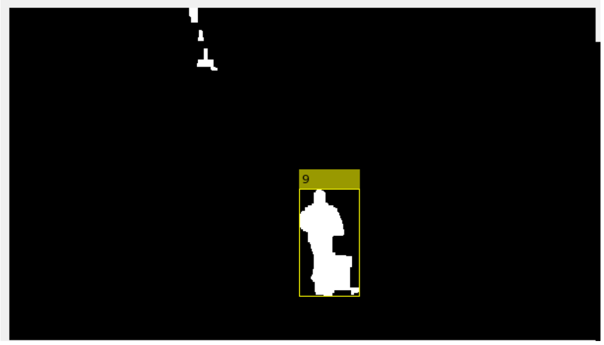

Figure 3: Multi-object Tracked CCTV Video Image

Then the performance of object detectors is utilized and carryout in the next section. In the experimentation work proposed and existing classifiers is implemented to five different CCTV footages files namely atrium, DBoverview, Avigilon, Kit Sample, games.

B. Performance Evaluation

The results of different multi-object tracking schemas are evaluated and experimented based on the parameters like Peak Signal to Noise Ratio (PSNR), Mean Square Error (MSE), Normalized Correlation Coefficient (NC) and Structural Similarity Index (SSIM). The mathematical representation for these performance parameters are as given in equation (26) to equation (30).

Peak Signal to Noise Ratio (PSNR)

The PSNR (τ_x) in dB is given as,

$$\tau_x = 10 \log_{10} \frac{R^2}{\mu_x} \tag{26}$$

Where, R is the maximum possible value in the related video frames and $\mu_x$is Mean Square Error (MSE) .

Mean Square Error (MSE)

The Mean Square Error (MSE) is given as,

$$MSE = \mu_x = \frac{1}{T}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(I_x(i,j) - I_x^{'}(i,j)\right)^2 \tag{27}$$

Where $I_x(i,j)$is the original video frames, $I_x^{'}(i,j)$is the noise removed video frames, and M and N are video frame height and width such that T=M×N. In this work, for video, PSNR is determined by taking average of PSNR values of all corresponding frames of the video. The average PSNR is determined as follows,

$$\bar{\tau} = \frac{1}{F}\sum_{x=1}^{F}\tau_x \tag{28}$$

Normalized Correlation Coefficient (NC)

The third parameter is Normalized Correlation coefficient (NC) is used for finding the similarities among original and noise removed video frames. The value of NC as ‘1’ indicates that the tracked multi object is highly correlated to that of the noise removed video frame and the value of NC as ‘0’ indicates that the tracked multi object is highly uncorrelated to that of the noise removed video frame. In general correlation value ranges is varied from 0 and 1. It is obvious from the correlation coefficient that more the value of NC, then the extracted watermark is closer towards the original. For each tracked video (corresponding to each frame of the video), the correlation coefficient (NC) is computed using the equation (28).

$$NC_x = \frac{1}{T}\sum_{i=1}^{M}\sum_{j=1}^{N}\overline{I(i,j) \oplus I'(i,j)} \tag{29}$$

Where, T=M×N represents the total number of pixels of object tracked video for xth frame. x varies from 1 to F. Exclusive- NOR operation is also performed here to obtain the NC value.

Structural Similarity Index (SSIM)

At finally compute a new parameter SSIM. It is used to measure and evaluate the similarity among noise removed video frame and original video frame,

$$SSIM(I,I^{'},x) = \frac{\left(2\mu_x\mu_y + c_1\right)\left(2\sigma_{xy} + c_2\right)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{30}$$

Where, I is the original video and I’ is the object tracked video from xth frame, μ_m is the mean of the intensities available in the original input video of xth frame, μ_n is the mean of the intensities presented in the object tracked video from xth frame. On the similar grounds, $\sigma_x^2$is variance of original video I, $\sigma_y^2$is the variance of object tracked video I$^{'}$ and σ_{xy} is covariance of original video and object tracked video of xth frame.

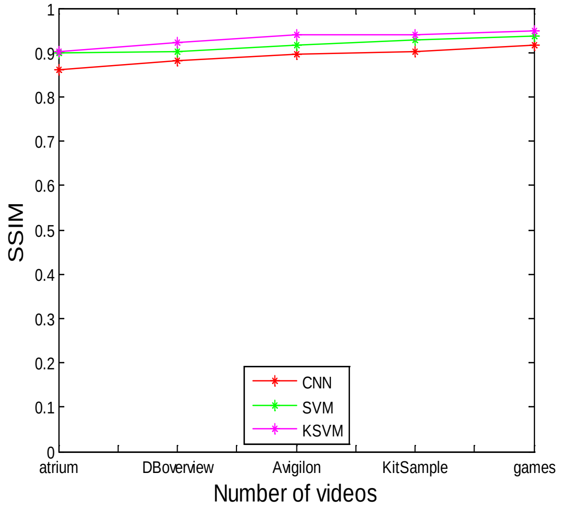

Figure 4: SSIM Comparison of Videos

Table 1: SSIM Comparison of Videos

S.No	*Videos*	*SSIM*		
		CNN	*SVM*	*KSVM*
1	atrium	0.8600	0.8981	0.9023
2	DBoverview	0.8821	0.9028	0.9228
3	Avigilon	0.8971	0.9179	0.9389
4	KitSample	0.9024	0.9281	0.9397
5	games	0.9181	0.9372	0.9483

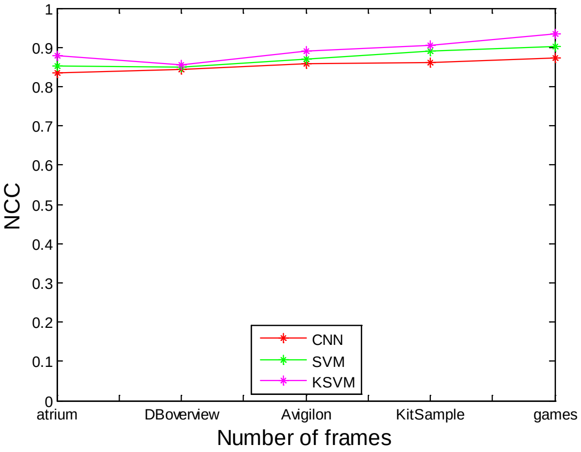

Figure 5: NCC Comparison of Videos

Table 2: NCC Comparison of Videos

S.No	*Videos*	*NCC*		
		CNN	*SVM*	*KSVM*
1	atrium	0.8362	0.8517	0.8798
2	DBoverview	0.8436	0.8489	0.8560
3	Avigilon	0.8596	0.8689	0.8893
4	KitSample	0.8618	0.8912	0.9036
5	games	0.8725	0.9018	0.9356

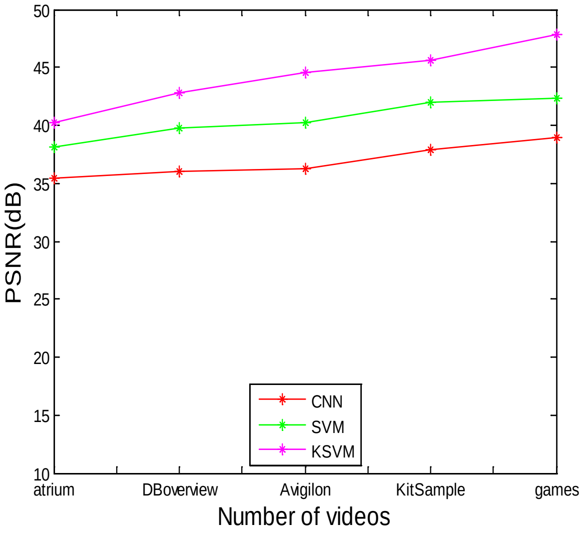

Figure 6: PSNR Comparison of Videos

Table 3: PSNR Comparison of Videos

S.No	*Videos*	*PSNR(dB)*		
		CNN	*SVM*	*KSVM*
1	atrium	35.43	38.17	40.23
2	DBoverview	35.97	39.73	42.86
3	Avigilon	36.23	40.18	44.58
4	KitSample	38.91	41.98	45.63
5	games	37.89	42.36	47.89

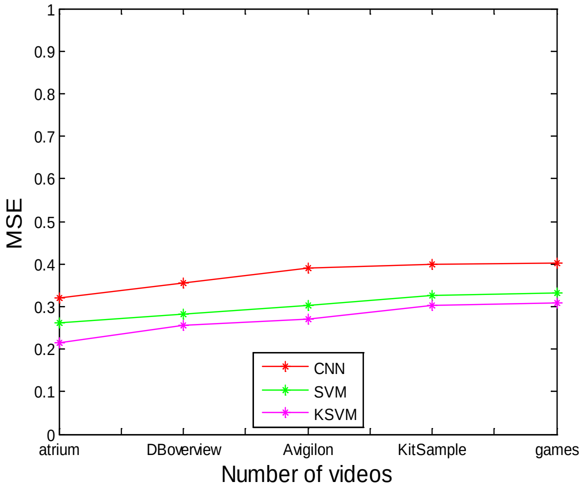

Figure 7: MSE Comparison of Videos

The graphs shown in Figure 4 to Figure 7, represents the Structural Similarity index (SSIM), Normalized Correlation Coefficient (NCC), Peak Signal to Noise Ratio (PSNR) and Mean Square Error (MSE) of the video and object tracked video under the condition is applied on the video. The graphs shown in all methods represents that the proposed

KSVM schema performs better for all parameters is applied on the CCTV video.

Table 4: MSE Comparison of Videos

S.No	*Videos*	*MSE*		
		CNN	*SVM*	*KSVM*
1	atrium	0.32	0.263	0.2153
2	DBoverview	0.356	0.2817	0.2563
3	Avigilon	0.3891	0.3025	0.2696
4	KitSample	0.3981	0.3258	0.3013
5	games	0.4018	0.3324	0.3089

The table 1 to table 4, represents the Structural Similarity index (SSIM), Normalized Correlation Coefficient (NCC), Peak Signal to Noise Ratio (PSNR) and Mean Square Error (MSE) of the video and object tracked video under the condition is applied on the video. The graphs shown in all methods represents that the proposed KSVM schema performs better for all parameters is applied on the CCTV video.

V. Conclusion and Future Work

In this paper a new Kernel Support Vector Machine (KSVM) based multi-target detectors is proposed to enhance the performance of multi-target detectors in real time CCTV footage. On the other hand the performance is reported in all original video frames, noised video frames and blurred video frames have been becomes a very challenging and difficult task in CCTV videos. To solve this problem, a novel Motion Adaptive Gaussian Filtering (MAGF) methodology is proposed to automatically remove noises from the CCTV footage at the time of concentrated PTZ operations. Moreover, graph partitioning based video segmentation algorithm partitions an input video into several frames. Each frame is heterogeneous with respect to one or more properties *i.e.* the variation of measurements within the regions should be considerably less than variation at the object borders. Moreover, a novel KSVM multiple objects tracking are proposed that automatically tune detector parameters at the time of intense PTZ operations. The KSVM based multiple object visual tracking for CCTV footage performs better than the conventional object tracking method ,since the proposed KSVM work is performed based on features and noises in the CCTV footage of the object is removed using filtering. In KSVM multiple object visual tracking schema, spatial transformation of the objects is used to train a KSVM to calculate the inherent camera characteristics in the next frame. The forecasted parameters are used toward adjust the multi –object detector parameters, foremost to improve the robust results. The experimentation results show that the dynamic scaling considerably enhances the accuracy of the object detector when compared to fixed scale operations. Semi-supervised methods have been used to discover noisy labels and adjust them so it is considered as scope of future work. Moreover, prior information regarding data source with the purpose of have more labeling noise than others have been used to find and adjust noisy labels. A further way of enhancing the performance of classifier is using hyper-parameter tuning with the purpose of identifying the greatest KSVM architecture by means of layer sizes and properties, and detecting the best quantity of video redundancy from tracking sources. Some more additional evaluation is done by considering negative or background class. Multi Object tracking task with KSVM have been implemented for the real time system.

References

[1] D. Dawson, P. Derby, A. Doyle, C. Fonio, L. Huey and M. Johnson, "A Report on Camera Surveillance in Canada Part two", Technical report, Social Sciences and Humanities Research Council of Canada, 2009.

[2] P. Kolekar, "Global and China surveillance cameras industry market research report". Technical report, Market Research, 2010.

[3] D. Wilson and A. Sutton, "Open-Street CCTV in Australia : A comparative study of establishment and operation", Technical Report April, University of Melbourne, 2003.

[4] A.C Davies and S.A Velastin, "Progress in computational intelligence to support CCTV surveillance systems", International Journal of Computing, Vol. 4, No. 3, Pp. 76–84, 2014.

[5] A.W. Smeulders, D.M. Chu, R. Cucchiara, S. Calderara, A. Dehghan and M. Shah, "Visual tracking: An experimental survey", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 36, No. 7, Pp. 1442–1468, 2014.

[6] H. Li, Y. Li and F. Porikli, Computer Vision-ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, November 1-5, 2014, Revised Selected Papers, Part V, chapter Robust Online Visual Tracking with a Single Convolutional Neural Network, Springer International Publishing, Cham, Pp. 194–209, , 2015.

[7] X. Zhou, L. Xie, P. Zhang and Y. Zhang, "An ensemble of deep neural networks for object tracking",IEEE International Conference on Image Processing (ICIP), Pp. 843–847, 2014.

[8] J. Ding, Y. Huang, W. Liu and K. Huang, "Severely blurred object tracking by learning deep image representations", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 26, No. 2, Pp. 319-331, 2016.

[9] J. Fan, W. Xu, Y. Wu and Y. Gong, "Human tracking using convolutional neural networks", IEEE Transactions on Neural Networks, Vol. 21, No. 10, Pp. 1610–1623, 2010.

[10] N. Wang and D.Y. Yeung, "Learning a deep compact image representation for visual tracking", in Advances in Neural Information Processing Systems, Pp. 809–817, 2013.

[11] L. Wang, T. Liu, G. Wang, K.L. Chan and Q. Yang, "Video tracking using learned hierarchical features", IEEE Transactions on Image Processing, Vol. 24, No. 4, Pp. 1424–1435, 2015.

[12] Y. Chen, X. Yang, B. Zhong, S. Pan, D. Chen and H. Zhang, "Cnntracker: Online discriminative object tracking via deep convolutional neural network", Applied Soft Computing, Vol. 38, Pp. 1088–1098, 2016.

[13] J. Weinman Jerod, Allen Hanson and Erik Learned-Miller, "Joint Feature Selection for Object Detection and Recognition", University of Massachusetts-Arhmerst Technical Report, 2006.

[14] C. Bregler, "Learning and Recognizing Human Dynamics in Video Sequences", IEEE Conference on Computer Vision and Pattern Recognition, Puerto Rico, Pp. 568-574, 1997.

[15] T.N. Tan, G.D. Sullivan and K.D. Baker, "Model Based Localization and Recognition of Road vehicles", International Journal in Computer Vision, Vol. 29, No. 1, Pp. 22–25, 1998.

[16] C.R. Wren, A. Azarbayejani, T. Darell and A.P. Pentland, "Pfinder: Real-Time Tracking of the Human Body", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.19, Pp. 780-785, 1997.

[17] S.M. Yoon and H. Kim, "Real-Time Multiple People Detection Using Skin Color, Motion and Appearance Information", International Workshop on Robot and Human Interactive Communication, Pp. 331-334, 2004.

[18] N. Dalal and B. Triggs, "Histogram of Oriented Gradients for Human Detection", Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Vol. 2, Pp. 886-893, 2005.

[19] B. Stanley and L. Wolf, "A Unified System for Object Detection, Texture Recognition, and Context Analysis Based on the Standard model Feature Set", Proc. British Machine Vision Conference, 2005.

[20] H. Schneiderman, "Learning a Restricted Bayesian Network for Object Detection", IEEE Conference on Computer Vision and Pattern Recognition, 2004.

[21] C.B. Alexander, T.L. Berg and J. Malik, "Shape Matching and Object Recognition using Low Distortion Correspondences", IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005.

[22] S.B. Kotsiatis, "Supervised Machine Learning: A Review of Classification Techniques", Informatica Vol. 31, Pp. 249-268, 2007.

[23] W. Hu, T. Tan, L. Wang and S. Maybank, "A Survey on 249 Visual Surveillance of Object Motion and Behaviors", IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, Vol. 34, No. 3, Pp.334-352, 2004.

[24] A. Yilmaz, J. Omar and S. Mubarak, "Object Tracking: A Survey", ACM Computing Surveys, Vol.38, No.4, 2006.

[25] C. Veenam, M. Reinders and E. Backer, "Resolving Motion Correspondence for Densely Moving Points", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 23, No.1, Pp. 54-72, 2001.

[26] C. Yunqiang, Y. Rui and T.S. Huang, "JPDAF Based HMM for Real-Time Contour Tracking IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, 2001.

[27] R. Karlson and F. Gustafsson, "Monte Carlo Data Association for Multiple Target Tracking", IEEE Target Tracking: Algorithms and Applications, Netherland, 2001.

[28] K. Rank, M. Lendl and R. Unbehauen, "Estimation of image noise variance", Proceedings: Vision, Image and Signal Processing, Vol. 146, No. 2, Pp. 80–84, 1999.

[29] M. Grundmann, V. Kwatra, M. Han and I. Essa. Efficient hierarchical graph-based video segmentation", Computer Vision and Pattern Recognition (CVPR), 2010.

[30] A. Vazquez-Reina, S. Avidan, H. Pfister and E. Miller, "Multiple hypothesis video segmentation from superpixel flows", European conference on Computer vision, 2010.

[31] A.Y.C. Chen and J.J. Corso, "Propagating multi-class pixel labels throughout video frames", Western NY Image Processing Workshop, 2010.

[32] C. Xu, S. Whitt and J.J. Corso, "Flattening supervoxel hierarchies by the uniform entropy

slice", Proceedings of the IEEE International Conference on Computer Vision, Pp. 2240-2247, 2013.

[33] G.J. Burghouts, A.W.M. Smeulders and J.M. Geuse broek, "The distribution family of similarity distances", Neural Information Processing Systems, 2007.

[34] C.P. Yu, W.Y. Hua, D. Samaras and G. Zelinsky, "Modeling clutter perception using parametric proto-object partitioning", Advances in Neural Information Processing Systems, Pp. 118-126, 2013.

[35] T. Steihaug, "The conjugate gradient method and trust regions in large scale optimization", SIAM Journal on Numerical Analysis, 1983.

[36] W. Zhang, S. Teng, H. Zhu, H. Du and X. Li, "Fuzzy Multi-Class Support Vector Machines for Cooperative Network Intrusion detection", Proc. 9th IEEE Int. Conference on Cognitive Informatics (ICCI'10), Pp. 811-818, 2010.

[37] S. Ahmad, A. Kalra and H. Stephen, "Estimating soil moisture using remote sensing data: A machine learning approach", Advances in Water Resources, Vol. 33, No. 1, Pp. 69–80, 2010.

[38] S. Yin, X. Xie, J. Lam, K. C. Cheung and H. Gao, "An improved incremental learning approach for kpi prognosis of dynamic fuel cell system", IEEE Transactions on Cybernetics, 2015.

[39] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn and A Zisserman, "The Pascal visual object classes (VOC) challenge", International Journal of Computer Vision, Vol. 88, No. 2, Pp. 303–338, 2010.

[40] A. Ess, B. Leibe, K. Schindler and L. Van Gool, "A mobile vision system for robust multi-person tracking", IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08), 2008.