# *Video Surveillance using Deep Learning - A Review*

Mrs. Shana L
*Department of Compter Science and Engineering*
St. Xavier's College of Engineering
Tamilnadu, India
shanbiji@gmail.com

Dr. C Seldev Christopher
*Department of Compter Science and Engineering*
St. Xavier's College of Engineering
Tamilnadu, India
cseldev@gmail.com

***Abstract*— Video surveillance is a rapidly growing industry. Video surveillance, more commonly called CCTV (closed-circuit television), is an industry that is more than 30 years old and one that has had its share of technology changes.**

**Video Surveillance has become an indispensable component for ensuring public safety in the modern world. Sophisticated video object tracking techniques specially designed for surveillance applications are of increasing importance for analyzing and understanding numerous surveillance videos in an effective manner. A large majority of video surveillance applications are concerned with monitoring activities within structured environments, such as indoor environments, surrounding areas of buildings, highways, traffic junctions, etc., the structures of which are often static and known to the surveillance personnel. One important characteristic of moving objects in these applications is that the motions of objects are constrained by the structure of the environment under surveillance. Therefore, it is beneficial and essential to explore the impacts of the environments upon the object motions, and integrate them into object tracking for improved performances.**

**The problem of video surveillance has been well studied which has been adapted for several issues. The behavior of any human can be monitored through video surveillance. There are number of approaches available for the video surveillance and behavior analysis. The previous methods uses background models, object tracking for the problem of behavior analysis.**

**Pervasive usage of video surveillance is rapidly increasing in developed countries. Continuous security threats to public safety demand use of such systems. Contemporary video surveillance systems offer advanced functionalities which threaten the privacy of those recorded in the video. There is a need to balance the usage of video surveillance against its negative impact on privacy.**

**Keywords— Video Surveillance, Object Tracking, Behavior Analysis, ANN**

## I. INTRODUCTION

The video surveillance is the process of monitoring the happenings in any indoor or outdoor location. Many organizations have engaged video surveillance systems for different purposes like security management, employee monitoring and so on. By adapting video surveillance systems, the administrator would be able to monitor the happenings and events in the location and can monitor the activities of any human. Similarly, the behavior of any human can be monitored and tracked using video surveillance systems. There are number of approaches available for video surveillance like object and template based models. The template models maintain list of templates for different activities based on that the activities has been performed. Similarly, in object based models, the method uses list of objects and their features in identifying the activity or behavior of the user. Similar to that, number of scientific approaches available for behavior analysis. The k-means algorithm would estimate the distance measure on different objects of specific behavior to identify the class of activity. The accuracy of the k-means algorithm is depending on the feature being considered and the number of samples available in each class. Also, the accuracy depends on the variant features considered. The support vector machine has been used for different problems and can be used for the task of behavior analysis in video surveillance. Whatever the algorithm being used, the accuracy is highly depending on the feature considered and the methods of similarity estimation.

The process of behavior analysis from video surveillance is starts with object tracking. The video would contain number of objects in the sequence of frames. In order to perform behavior tracking, it is necessary to identify the objects present in the video frames. Once the objects of the frame have been detected then behavior can be identified. The objects can be identified by maintaining the templates of the objects of various activities. The template matching can be performed to identify the objects. Similarly, the background features plays vital role in behavior analysis. By removing the background features, the objects can be identified efficiently. The background features has been eliminated by using different background models but produces poor accuracy in removing the background features.

By considering all these, a multi variant feature analysis model has been presented in this paper. The method uses multiple features of variant form to perform behavior analysis model. The artificial neural network has been used for classification. The ANN has great impact in various scientific problems and the same can be used for behavior analysis in video surveillance.

## II. LITERATURE SURVEY

The proposed methodology is segmented into three phases such as detection phase, tracking phase and evaluation phase. This proposed can be used in various environments for the detection and tracking of human. The human detection and tracking consist of four phases such as

1. Preprocessing
2. Foreground Segmentation
3. Detection and Tracking Phase

978-1-5386-9482-4/19/$31.00 ©2019 IEEE

4. Evaluation Phase

A. Preprocessing

Noise removal by fuzzy morphological filter

In [2], the main idea of fuzzy morphological filter is the selection of Structuring Elements (SEs). Image demonizing effect is done by different structural elements. Morphology is a mathematical framework for the analysis of spatial structures. It is based on set theory. (Mahalingam et al., 2018).

In [3], a new fuzzy filter is presented which remove the random impulse noise (Melange et al., 2011). In this a filter is presented which prevent the filtering of noise free image pixels. This work can be extended towards the denoising of video sequences contaminated with fixed valued impulse noise and towards color image sequences.

In video surveillance for the input data an observation is generated for every frame. The observation includes the information that are used for the pose tracking methodologies. A back ground subtraction method is needed to extract the human body silhouette from every view. The contrast background is removed from every image and the result is the body including the colour image. (Alexandros et al. 2015)

The pre-processing module consists of two sub process: Filtering and Threshold Process. The filter Prewitt operator is used to remove some features present in the image. The Prewitt operator is calculated with kernel Kx and the horizontal edge component is calculated with kernel Ky, |Kx|+|Ky|. This indicates the gradient of current pixel. Thresholding is based on threshold value which is used to convert gray scale image into binary image. The main challenge is to find the threshold Value. Adaptive threshold is the method in which the different threshold value is used for different regions. (C S Sanoj et al. 2013)

B. Foreground Segmentation

The foreground detection is done by background subtraction. The background subtraction is done by the threshold difference of the current image and reference image [1]. A mixture of Gaussian, Nonparametric Kernal and codebook is used which have better performance. The main drawback of this method is that this method needs extra expensive computation and more memories. [Jianpeng Zhou & Jack Hoang, 2005].

One of the commonly used methods for segmenting an image is based on the histogram based segmentation. In [8], the histogram of the distance is measured by the peaks obtained in the plot. The leverage of these peaks can be reduced by dividing the images into different sub images. For better segmentation, the first average distance is measured by Haar cascade based detection and the objects behind the target object are eliminated. This method is concentrated for the detection of upper body part and face, since they are visible without any occlusion. The two areas are different when they are wrapped with a rectangle window. The area above the shoulder is unstable so that it badly influenced if the disturbances is not eliminated.

C. Feature Selection

Feature selection and extraction is an essential and key step in human detection. Unique and discriminative features are extracted from the input scene for human detection. The tracking of non stationary objects from the video frames is done in the detection phase. For the detection and tracking of human, the following features are extracted.

HOG Features

In this method the video is divided into frame. Each video frame is divided into small regions and the histogram gradient directions and edge orientations are computed over the block of pixel. An extended version of this algorithm can be used for the problems related to illumination and shadowing.

Histogram of the gradient is a feature descriptor, which is most commonly used for object recognition. The vertical and horizontal gradients are calculated by the algorithm. In[5], histogram is used for the extraction of patterns. Patches in multiple images is analyzed by the gradient histogram algorithm. The patches can be cropped and resized. The absolute value of the x and y gradient is used to calculate to extract the features.

Haar Features:

Haar feature is most used to find the structural similarities between samples of the class of humans. The Haar Wavelets includes the functions which capture change in intensity of the image along all directions. The Haar functions change the intensity in horizontal, vertical and diagonal directions. Improved version of the method is applied to multi scale exposure of pedestrian .

Haar features have high scalar values that represent average intensities between the regions. They capture the gradient of the intensity at different locations, frequencies and directions. The Haar basis functions extract the important features of the image. By extracting the features the performance of the classifier is enhanced.

Texture:

An image texture is a set of metrics calculated in image processing designed to quantify the perceived texture of an image. By considering the textual features the classification of pedestrian is only a challenging problem due to the high variability. Texture feature is used in conditions like varying light condition and variations due to clothing. This feature is normally used along with features such as shape, colour and others.

LBP [13] means Local Binary Pattern describes images based on their texture. LBP characterize the image by considering the neighborhood of each pixel. This method is robust against variations in pose or illumination than other methods. So this method is widely used.

BWT and ZM:

Several methods have been proposed that use a single feature for human detection. It is difficult to represent different types of objects, shape, actions etc using a single feature. In [4], a suitable strategy synthesizing more than one

feature type is used. One feature is suitable to represent human and other is used to detect other type of objects. The methods BWT (Bi-orthogonal Wavelet Transforms) and ZM (Zernike moments) adopted the combination of two features.

BWT is of two types Linear Phase BWT and Shift Invariance BWT. ZM represents the properties of an image with no redundancy or overlap of the information between the moments. This is applicable to different problems such as object classification, shape analysis, content based image retrieval etc. The properties of ZM are invariance, robustness, expressiveness, multilevel representation, effectiveness.

Local Distance Features (LDF)

In [15] this paper the depth maps is converted into binary edge contour to develop human shape edges bounded by data. In this ten different joints such as head, torso, left/right elbow, left/right hand, left/right knee and left/right foot. The Euclidean distance between the initial joint is measured and then for the next joint is also measured. This process is measured.

D. Human Detection and Tracking Phase - Machine learning Techniques

The following are some of the machine learning techniques used so far.

SVM – Support Vector Machine

Support Vector Machine is used as a classification method. Support Vector Machine improves the efficiency of machine learning. It works on the principle of fitting a boundary to a region of points which are all belong to one class. SVM has given a good outcome for data classification. In [5] based on the SVM classification of individual objects, each action will be recognized independently.

In [7], the support vector machine classifier is used as a two classifier. The SVM classifies the cropped images in the dataset as positive samples and negative samples. The classification accuracy is very high in the order of 98.32%.

Artificial Neural Network (ANN)

Neural networks are parallel computing devices, which are basically an attempt to make a computer model of the brain. The main objective is to develop a system to perform various computational tasks faster than the traditional systems. Artificial Neural Networks (ANNs) performances is up to the benchmark databases. A stereo-system for pedestrian detection and classification method is developed by Zhao et al.

The ANN has great impact in various scientific problems and the same can be used for behavior analysis in video surveillance.

Deep CNN Learning

Deep learning is getting a lot attention for good reasons. It is achieving unprecedented levels of accuracy. Deep learning method is used in many fields like automatic object detection, Image segmentation and Classification due to its accuracy and less computation time.

In [6], UAV Videos are becoming prominent and used widely for surveillance and reconnaissance. In this paper a actively detect, track and classify multiple systems is represented. In addition to this additional information such as coordinates and velocities of detected objects can also be calculated. The detected object is classified as a vehicles or humans by pre-classification and deep convolution neural network. The vehicles and humans are highly separable due to their huge size. Due to this an adaptive threshold is used to detect the size. The Deep CNN is used because of its favorable performance in image classification.

E. Behavior Analysis

The problem of behavior analysis has been approached with different methods. Such methods are discussed and reviewed in this section.

In [17], the author presented a detailed review in behavior understanding in surveillance videos. The author details various human activities and how it can be identified through surveillance video. Also, different methods have been studied and compared their performance methods and measures in detail.

In [18], the object tracking has been performed with Bayesian model. The method uses distance measures of objects in a structured environment. The method also used the packet filtering in object tracking.

To track multiple objects in surveillance video in online stream, a Bayesian model with top down approach is presented in [19]. The method focused on labeling multiple objects in video stream. The method adapts the occlusion features and finite sets of the multiple objects. The Bayesian model is approached recursively to track the objects of the video.

The correlation and dependency measures on feature for efficient multi label classification are presented in [20]. It is a Bayesian model which learns multiple class data streams. It measures the correlation between the streams and measures the relationship between them. According to the correlation measure and dependency measures, the classification is performed.

An HMM based behavior tracking model is presented in [21]. The model is a bottom up approach which transforms the image features into vector sequences. The vector sequences generated has been used to quantize the samples to perform classification on behaviors.

In DFP-ALC [22], the method extracts the patches of distinct frames. Such patches are indexed based on appearance based clustering (ALC). In classification, exact class is selected based on appearance measures.

Depth matrix based gesture recognition is presented in [23], which extract the image features and convert them into shape matrix from the posture sequences. Using the sequences, he method extracts the regional descriptors and classify them using Naïve Bayes Classifier.

The problem of recognizing the behavior of pedestrian, an reciprocal model to guide the pedestrian is presented in [24]. The method uses image and attribute features to measure the distribution weight. Estimated weight measure has been used to perform classification.

The behavior tracking has been approached with skeleton features in [25]. The method extracts the skeleton prototypes based on the shape features. Various pose of skeleton features are extracted and has been used to perform classification.

The problem of assessing autism spectrum disorder is handled with the pattern of robot and human interaction is presented in [26]. The method captures the behavior of children through RGB-D sensors. The method measure the behavior of the children based on the movements of body and head. Also the features of direction, magnitude of gazing's and energy of kinetic are used for classification.

In [27], the method segments the parametric and actions of primitives. The method combines the actions in to number of sequences and based on the similarity of motion sequences the classification is performed.

A knowledge based approach for controlling video surveillance has been presented in [28]. The model has leaned features of classes according to the features considered. The method performs classification according to the features available.

## F. CONCLUSION

This paper presented a detailed review on the methods used for pedestrian detection, tracking and behavior analysis. Recent Deep Learning methodologies for the detection of human and tracking deserved a dedicated state-of-the-art survey. The problem of behavior analysis has been approached with different methods. The reviewed works highlights the need to investigate new methods for the detect pedestrian, tracking and behavior analysis. The Deep Learning together with classical Machine Learning models has high levels of accuracy. This method requires less computation time with respect to the exixting features and classification.

## REFERENCES


[1] Jianpeng Zhou & Jack Hoang, "Real time robust detection and tracking system", 2005

[2] T Mahalingam, M Subramoniam, "A robust single and multiple moving object detection, tracking and classification", 2018.

[3] Shusheng HE, Alei Liang, Ling Lin and Toa Song, "A continuously adaptive template matching algorithm for human tracking", 2017.

[4] Om Prakash, Jeonghwan Gwak, Manish Khare, Ashish Khare & Moogue Jeon, "Human Detection in complex real scene based on combination of biorthogonal wavelet transform Zernike moments", 2018.

[5] Seemanthini K & Manjunath S S, "Human Detection and Tracking using HOG for Action Recognition", 2018.

[6] Husetin Can Baykara, "Real Time Detection Tracking and Classification of Multiple Moving Objects in UAV Videos", 2018.

[7] Vandit Gajjar, Ayesha Gurnani & Yash Khandhediya, "Human Detection and Tracking for Video Surveillance: A Cognitive Science Approach", 2017.

[8] Sangeetha D & Deepa P, "Efficient Scale Invariant Human Detection using Histogram of Orientated Gradients for IoT Services", 2017.

[9] Shusheng He, Alei Liang, Ling Lin & Tao Song, "A Continuously Adaptive Template Matching Algorithm for Human Tracking", 2017.

[10] Yepeng Guan & Yizhen Huang, "Multi-pose human head detection and tracking boosted by efficient human head validation using ellipse detection", 2015.

[11] Ning HE, Jiaheng & Lin Song, "Scale Space Histogram of Oriented Gradients for Human Detection", 2008.

[12] Daegon Kim & Sung Chun Lee, "Pairwise Threshold for Gaussian Mixture Classification and its Application on Human Tracking Enhancement", IEEE, pp. 368-372.

[13] X. Wang, T.X. Han, S. Yan, "An HOG-LBP human detector with partial occlusion handling, in: Proceedings of the IEEE 12th International Conference on Computer Vision", IEEE, 2009.

[14] Melange, T, Nachtegael , M, Schulte, S & Kerre, E E, 2011, "A fuzzy filter for the removal of random impulse noise in image sequences", Image and Vision Computing, vol. 29, no.6, pp. 407–419.

[15] Ahmad Jalal, Shaharyar Kamal & Daijin Kim, "A Spatiotemporal Motion Variation Features Extraction Approach for Human Tracking and Pose based Action Recognition", IEEE 2013.

[16] Alexandros Moutzouris, Jesus Martinez-del-Rincon, Jean Christophe Nebel & Dimitrios Makris, " Efficient Tracking of Human Pose using a Manifold Hierarchy", 2015.

[17] C S Sanoj, N Vijayaraj & D Rajalakshmi, " Vision Approach of Human Detection and Tracking using Focus Tracking Analysis", IEEE, 2013. SarveshVishwakarma, AnupamAgrawal, A Survey on Activity Recognition and Behavior Understanding in Video Surveillance, Springer, Visual Computing, 2012.

[18] Junda Zhu, Object Tracking in Structured Environments for Video Surveillance Applications, IEEE transactions on circuits and systems for video technology, Vol 20, No 2, February 2010

[19] Du YongKim, A labeled random finite set online multi-object tracker for video data, Elsevier, Pattern Recognition, Volume 90, 2019.

[20] TienThanhNguyen, Multi-label classification via label correlation and first order feature dependance in a data stream, Elsevier, Pattern Recognition, Volume 90, 2019.

[21] Yamato, Junji,Recognizing Human Behavior Using Hidden Markov Models, Springer, Video Computing, 2002.

[22] SivapriyaaKannappan, DFP-ALC: Automatic video summarization using Distinct Frame Patch index and Appearance based Linear Clustering, Elsevier, Pattern Recognition, Vol.120, pp:8-16, 2019.

[23] LalitKanea, PriteeKhanna, Depth matrix and adaptive Bayes classifier based dynamic hand gesture recognition, Elsevier, Pattern Recognition, vol. 120, 2019.

[24] ZhongJi,Image-attribute reciprocally guided attention network for pedestrian attribute recognition, Elsevier, Pattern Recognition, vol 120,2019.

[25] AlessiaSaggese, Learning skeleton representations for human action recognition, Elsevier, Pattern Recognition, vol. 118, 2019.

[26] Salvatore MariaAnzalone, Quantifying patterns of joint attention during human-robot interactions: An application for autism spectrum disorder assessment, Elsevier, Pattern Recognition, vol. 118, 2019.

[27] Juan PedroBander, A new paradigm for autonomous human motion description and evaluation: Application to the Get Up& Go test use case, Elsevier, Pattern Recognition, Vol. 118, 2019.

[28] B. Georis, F. Brémond, M. Thonnat, Real-time control of video surveillance systems with program supervision techniques, Springer, Machine Vision and Applications, Volume 18, Issue 3–4, pp 189–205, 2007.