*Intelligent Systems and Computer Technology*
*D.J. Hemanth et al. (Eds.)*


*doi:10.3233/APC200175*

# A Review on Video Tracking for Multiple Objects in Neural Network

UmaS[a,1], and Umamaheswari S[b]
[a]*Research Scholar, Anna University, Chennai, India*
[b]*Assistant Prof., Anna University, Chennai, India*

**Abstract.** Video Multiple Object Detection and Tracking (VMODT) is the current area of research in computer vision which has increased due to the attention of commercial and academic potential it has offered. Multiple Object Tracking (MOT) shares all challenges to be handled such that long time occluded,fully occluded in small object, frequent occlusion in crowd and severely blurred in fast motion in video etc.,In this paper, a review on VMODT with various models,approaches and tracking algorithms is carried out.Neural Networks are developed to provide optimal solution in monitoring the regularity,re-identification,predicting the activity and control action of objects in surveillance application. This review work will be helpful for understanding the start-of-art in VMODT, finding the limitation in current algorithm. However, hybriding the different tracking methods will improve the performance and facilitate new approaches in deep learning.

**Keywords.**Camera Surveillance, Multiple objects detection, Tracking, Occlusion, Deep learning

## 1. Introduction

In computer vision, tracking is the main role in video analysis. Tracking of object needs detection of the object first. The object detection is the process of locating the target object in a frame with bounding boxes. Tracking is the process of finding the location of object in the first frame and then link these targeted object at all frames in video.The collection of sequence of framesin video.VMODT is the analysis ofmultiple objects tracking which includes main components as observation and tracking. Observation is the identification of target object in order to differentiate from multiple objects which have special and unique features like size, color, shape, motion etc., Motion and appearance are two main components which are independent in tracking process.From image, object is extracted using segmentation in first step. In second step, feature extraction is done on color, shape, texture pattern and finally object is classified.. In graph based model, the energy function is developed by using image appearance and its gradient. Filtering and clustering are the two steps in segmentation. The color, texture, moment features are used to find the class label of unknown object using KNN (K-Nearest Neighbour) clustering algorithm [2].

[1]Uma.S[a,]Associate Professor, Panimalar Engineering College, Chennai, India;
E-mail: umaokj@gmail.com

Deep Learning (DL) [9] in VMODT is robust due to the massive parallel data processes, labeling and by deep new architecture. Deep networks have many layers in network.Deep Learning approaches in architectureare Region Proposals (R-CNN, Fast R-CNN, Faster R-CNN), Single Shot MultiBox Detector (SSD), You Only Look Once (YOLO) [6].

The CNN [1] takes the current frame for reference whenever new frame enters into it. The selective search (RegionProposal) method is generally used to find location of target object in CNN, it grouping the similar regions based on size, shape, color, texture. In the first frame the object is detected using region based, then in DNN object is detected till end of the frame.The CNN and F-RCNN [1] are the concepts in deep learning to identify the behavioral metrics. The end-to-end deep learning architecture is designed to extract the motion information and appearance descriptor. The advantage of CNN is to exploit the local structure and internal geometric layout of the target.CNN is easier to train than the feed forward neural network. CNN supportsweight sharing, built in kernel,maxpooling, padding and stride which helps to reduce the parameter size of the image.

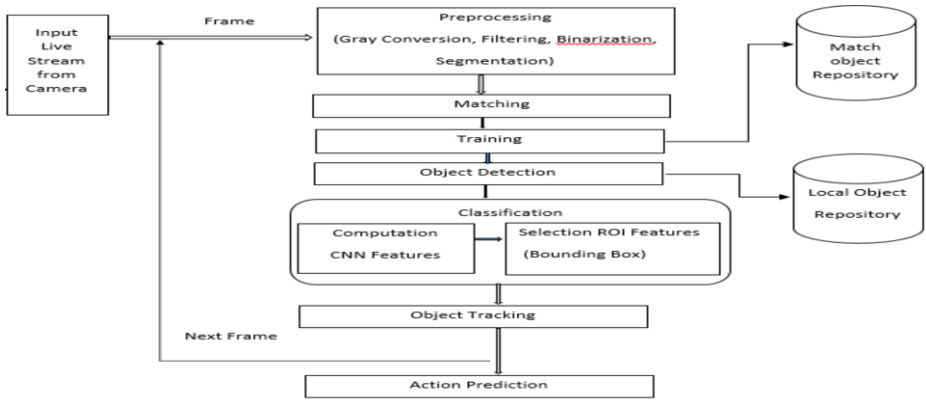


**Figure 1. Framework object tracking in neural network**

From Fig.1, the first frame from video is preprocessed that includes the gray color conversion, filtering noise, binarization for digital and segmentation of region. Then, the region of image is compared with match object repository. After training with data set, the object is detected and compared with local object repository. CNN features and ROI features are used to classify the objects. Next frame is processed and tracking is done till the end of the frame in video. The action of target object is predicted with the help of deep learning.

## 2. Literature Survey

Jakecowton et.al [1] proposed Convolution Neural Network (CNN),Faster Region-based Convolution Neural Network (F-RCNN) are the two tracking methods in real time video to pertaining the individual object behaviours.It extracts behavioral metrics particularly average speed, total distance travelled, spentthe idle time. The CNN and F-RCNN helps in detecting the health conditions of pig and wellbeing.

Y H Sharathkumar et.al [2] proposed the K-Nearest Neighbor method to classify the animal and Graph cut method to remove background clutter from image. The image is divided into blocks. The color texture moments features are retrieved from blocks. In this paper, the accuracy of object classification are reported as minimum, maximum and average level. The color distribution captured from lower-order moments. The local Fourier transform is used to extract the features. The KNN chooses the K-points from n-dimensional inputs. The distance between the test sample and train sample is done by using Euclidean distance method.The PNN (Probabilistic Neural Network) is the other approach to classify the animals but it takes more memory space and is slow in execution than KNN method.

Xiao Liu et.al [3], proposed a discriminative model to predict the object from labeled video. The target state is inferred using current time stamp and join probability score. The detection algorithm need post processing to choose some subset from total responses in detection as output, selection mistakes included.The joinprobability score and suppression is helpfulin reducing the risk in post processingmistake and improve the performance in tracking. Cutting plane optimization, convex optimization makes the performance efficiently in crowd and congested area surveillance. The Gaussian Mixture Model (GMM) is used in tracking the video sequence when multiple objects are crowded. The Monte Carlo (MC) method is proposed to track the object sequentially. The current observation together with past frame which is inferred for prediction in next frame in the real time video. The post processing is required to detect the multiple object. Prediction and updating are two processes in each frame. Max margin model is used to label the object. The Markov chain (first order) moment is used to track the last frame with observation from current frame.

ShuaiZhong et.al [4],proposed Mean Shift (MS) segmentation method to achieve detection and tracking in multiple camera which is not overlapped. Bayesian Kalman filter is better than the Gaussian Mixture to represent the state of object and noise. Blurred object is tracked with robustness in video. The object segmentation is done by using MS. To filter the region in object, Adaptive GMM (AGMM) is used. The difference between foreground pixel and total pixel exceeds the threshold, then the region of pixel belongs to foreground. The object identified using SP-EMD (Super pixel Earth Movers Distance) algorithm gives the shape and color features. The occluded object is segmented using Deep Map (DM) and K-means Clustering.

Schichao Zhao et.al [5], proposed Deep ConvolutionNet to provide better performance in classification task. Image ConvNets are helpfulto capture the action details in video.ConvNet is used to utilize temporal information of video and to encode video sequence. To represent the action on object the SpatialNet and TemporalNet is utilized. Trajectory pooling and line pooling are thetwo strategies in pooling. The pooling is encoded for representation of video by VLAD (Vector of Locally Aggregated Descriptors). Fishers Vector encode helps to encode the features based on first order, second order parameters.To compute the training set on pooled descriptor-GMM(Gaussian Mixture Model) is used.While analyzing the video, precomputing process is skipped by pooling of convolution layers

Joseph Redmonet.al [6] proposed a YOLO (You Only Look Once) concept in order to detect the object. In YOLO, it predicts the multiple bounding box and its

class probability simultaneously in a single convolution network. The entire frame is predicted using a single neural network with bounding box in single evaluation. Loss function is trained to improve performance in end to end detection.In YOLO 45 frames are processed per second whereas 155 frames are processe`din fast YOLO in real time.The DPM (Deformable part models),R-CNN are used to detect the objects from artwork frame. In an image, initially the bounding box is generated using region based methods using R-CNN and then the objects are classified.To eliminate the duplication in object detection post processing is used in YOLO.

Jifeng Dai et.al [7] proposed a region based method for accurate and efficient object detection in a fully convolution network. The region based computation is shared by the entire image. Position sensitive score map addresses the translation variation, translation invariance in detection and classification. The residual network method is the backbone for object detection which is adapted to classify the image. In R-CNN, the computation is done from cropped region. ResNet is faster, in training and inference. The extension of F-RCNN is developed for segmentation and detection.Fast R-CNN, Faster R-CNN, SPPnet are do the computation on entire image.

ShaoqingRen et.al [8] proposed the RPN (Region Proposal Network) for object detection. The next method SPPNet, Fast-RCNN are used to reduce the running time.RPN is trained to predict the objectness score and object bounds at each position. By combining the Fast R-CNN and RPN, a unified single network is defined. RPN is cost free by sharing convolution features. To improve quality and accuracy in object detection RBN is used.

Ahemed Ali Hammam et.al [9] proposed classic method to monitor and track the pet animals.IoT (Internet of Things) used to provide the interaction and control among human and pets. This paper, approached deep learning to detect the pet animals and classify it on video sequence. The various tracking devices like Smartphone, Camera, Tag, GPS and RFID are combined together through IoT to develop the business objectives widely. The optical flow method is used to implement the tracking. To localize the object, pre-trained approach is used in Fast R-CNN. Feature extraction of pet animals is done by Fisher Locality Preserving Projection algorithm. The face image of pets is classified with help of SVM (Support Vector Machine). Pet animals are detected and classified using Fast R-CNN. The cat pet animal action is tested in this model with various video frames.

AlexKrizhevsky et.al [10] proposed deep CNN to classify the images as classes using supervised learning. Neural network have convolution layers which have millions of parameters and neurons with softmax activation function and fully convolution layer. Convolution in GPU and non-saturating neurons helps to make training faster. Dropout method is developed to solve overfit problem in fully connected layers. To provide the good results the depth in number of layers in convolution layer is required in video sequence. The memory in GPU and training time limits the network size. Fishers Vectors (FVs) helps to the prediction of classifiers. Learning is trained by stochastic gradient descent.

Junwei Li et.al [11] proposed a novel method for online tracking to display the appearance of the target object. Tracking performance was improved by encoding target appearance using crafted features and structured output learning.

Convolution features of the target are extracted from the first frame by kernels. A new strategy is proposed to capture the variation in the appearance during tracking to update the target and background kernel pool .To reduce the uncertainty in labeling the samples target location is refined using structured output SVM.

S J Sugumar et.al [12], proposed the EIDS (Elephant Image Detection System) with unsupervised learning method. This EIDS which solve the human-elephant conflict problems such as crop damage, human killed elephant and vice versa. In the forest area, ifthe elephant image is once identified, it is sent to base station EIDS system. By using image vision algorithm the image are decomposed and feature extracted to track it. The Euclidean and Manhattan distance method is enforced to track the elephant. Finally, the alert message is sent to forest officials by GSM.

YingkunXu,et.al [13], proposed deep learning to track multiple objects to solve confusing appearance,occlusion, in-and-out objectsand inadequacy of labelled data problems. MOT2015, MOT2016 are the two datasets used to detect the pedestrian. PET 2009, ETHMS ,KITTI dataset are used for tracking. End-to-end deep learning architecture is designed to extract the features for appearance, motion information. The CNNTis used to find the distance metric between detection and trackers. CLEAR and VACE metrics are used to evaluate the performance of MOT algorithms.

S Malathi, et.al [14], used Cauchy distribution model to compare current frame with previous frame in video. If any changes in environment, the Absolute Differential Estimation method is used to detect the object. A motion is detected, when the threshold value is exceed among the reference frame and current frame. GCM (Google cloud messaging) is used to send the notification to the user regarding the target object. The virtual panic button is also used for indication to the authorized department.

## 3. Conclusion

In this paper, we reviewed VMODT usingvarious tracking approaches such as DNN, R-CNN,Fast R-CNN. The different algorithms in tracking like kalman filter, Mean shift, Particle filter, Gaussian Mixture , Markov chain and Monte Carlo are reviewed for detection and tracking in video. Neural network have been promoted to provide the optimal solution in monitoring, re-identification, prediction and controlling the action in various surveillance application. Additionally if the classes of animals are similar, then there is a need for a more efficient tracking algorithm. This review work will be helpful for a basic understanding in VMODT, limitation in various algorithms in tracking and to facilitate the new approaches like graph model and network flow by deep architectures.

## References

[1] Jake cowton,IliasKyriazakiset.al:Automated Individual Pig Localization ,Tracking and Behavior Metric ExtractionUsing Deep Learning.IEEE ACCESS,2019,7,pp.108049-108060

[2] Y H SharathKumar,Manohr N &chetan N K .Animal Classification System; A Block based Approach. Internationalconference on advanced Computing Technologies and Applications, Elsevier,2015,pp. 336-343]

[3] Xiao Liu,Dacheng Tao et.al.Learning to Track Multiple Targets.IEEE Transaction on Neural networks and Learning system, 2015, 26,(5), pp.1060-1073

[4] ShuaiZhang,Chongwang et.al: New Object Detection ,Tracking, and Recognition Approaches for video surveillanceOver Camera Network .IEEE SENSOR JOURNAL,2015 ,15,(5),PP. 2679-2691

[5] Shichao Zhao, YanbinLiu,et.al: Pooling the Convolutional Layers in Deep ConvNets for Action Recognition .IEEE Transactions on Circuits and Systems for Video Technology,2015, pp.(99).1-11

[6] Joseph Redmon, Santosh Divvala ,et.al: .You Only Look Once: Unified, Real-Time Object Detection.Computervision foundation,IEEE Explorer,2018,pp. 779-788

[7] JifengDai,Yi Li et.al:R-FCN:object Detection via Region-based Fully Convolutional“, Neural information processing NIPS,2016,pp. 1-9

[8] ShaoqingRen, KaimingHe,et.al:”Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.IEEEAND MACHINE NTELLIGENCE, 39, (6),2017,pp.137-1149

[9] Ahmed Ali Hammam,MonaM.Solliman et.al:” Deep Pet:a Pet Animal Tracking System in Internet of Things using Deep Neural Network”IEEE,2018,pp. 38-43

[10] Alex Krizhevsky, IlyaSutskever ,et.al:“ImageNet Classification with Deep Convolutional Neural Networks.COMMUNICATIONS OF THE ACM,2017,60 ,(6),pp. 84-90

[11] JunweiLi,Xiaolong Zhou et.al: Object tracking using a convolution network and a structure output SVM.computationvisual media,2017,3,(4),pp.325-335

[12] S.J. sugumar and R.jayaparvathy.An improved real time Image detection system for Elephant Intrusion along the ForestBorder Areas .The scientific world channel,2014,

[13] YingkunXu, Xiaolong Zhou, shengyongchen, &Fenfen Li :"Deep learning for multiple object tracking :a Survey .IET Computer vision, 2019, 13, (4) , pp. 355-368

[14] Dr S.Malathi,et.al.Smart Video Surveillance System and Alert with Image Capturing using Android Smart Phone.2014, IEEE, International Conference on Circuit, Power and Computing Technologies,pp 1714-1722