ISBN 978-1-6654-3986-2

MULTI-OBJECT RECOGNITION AND TRACKING WITH AUTOMATED IMAGE ANNOTATION FOR BIG DATA BASED VIDEO SURVEILLANCE

K. Vijiyakumar[1], V. Govindasamy[2], V.Akila[3]
[1]*Research Scholar, Department of Computer Science and Engineering, Pondicherry Engineering College, Pondicherry, India. E.mail. vijiya.kumar@gmail.com*
[2]*Associate Professor, Department of Information Technology, Pondicherry Engineering College Pondicherry, India. E-mail. vgopu@pec.edu*
[3]*Assistant Professor, Department of Computer Science and Engineering, Pondicherry Engineering College ,Pondicherry, India. E-mail.akila@pec.edu*

2021 International Conference on System, Computation, Automation and Networking (ICSCAN) | 978-1-6654-3986-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICSCAN53069.2021.9526403

Abstract-Presently, the scope and application of Big Data Analytics in video surveillance makes it possible in different domains. In the area of intelligent visual surveillance, the procedure of tracking is described as finding a path or trajectory of an object of a given video sequence. Multi-Object tracking (MOT) mechanism become more familiar because of its applicability in numerous ways. Generally, MOT is employed to predict the position of various specified objects across multiple consequent frames with the offered ground truth position of the target in the beginning frame. In this paper, we have introduced an improved region based scalable convolution neural network (IRS-CNN) based MOT model. The presented IRS-CNN model enhances the existing RS-CNN by incorporating an automated image annotation (AIA) tool for increasing the detection rate as well as reducing the computation time. The interesting feature of AIA tool helps to rapidly annotate the training images in an automatic way. The novel IRS-CNN approach is tested against a benchmark UCSD anomaly detection dataset. A broad experimental result verified the optimal behavior of IRS-CNN model against a set of applied test images over the compared methods.

Keywords: Computer vision; R-CNN; MOT; Recognition; Image Annotation

I. INTRODUCTION

A smart surveillance system gained more research interest in the area of computer vision. Regularly, massive amount of videos has been recorded by numerous cameras deployed in a city for various surveillance purposes. In recent days, the requirement for smart visual surveillance has been increased because of the increasing significance of public security at diverse places.

In video surveillance applications, multiple objects tracking is an essential issue in videos that comprise extensive applications in video analysis scenes like sports analysis, autonomous driving, visual surveillance and robot navigation. To support tracking when certain type of object needs to be tracked like cars or people, some kind detector might be used. Modern development over Multi Object Tracking (MOT) had aimed over tracking through detection method from an object detector; object detections are associated to formulate goal trajectories. To determine the ambiguities in linking object detections and to avoid failure of detection, video series are processed in a batch mode [2]. The overall operation involved in MOT process is displayed in Fig. 1.

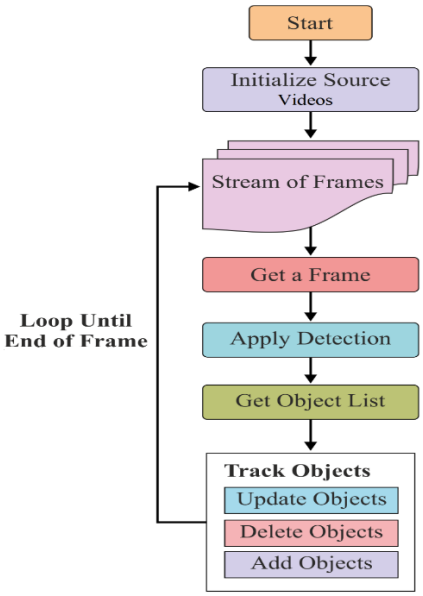

Fig. 1. Overall principle involved in MOT

The main difficulty for online tracking-by-detection is how to link with the noisy object detection in present video frame with formerly tracked objects. The foundation for the data association technique is a similarity function among targets and object detection. It is highly helpful to merge various computing cues like motion, location and appearance to manage the ambiguities through linkage. To adjust the attributes through cross-validation and to select parametric models heuristically for similarity function, it is not easily scalable to the feature count and it does not make sure model generalization power necessarily.

There exists a tendency over learning towards track which advocates the injecting learning concept abilities toward MOT [3,4]. We divide the techniques of MOT towards online and offline learning techniques. Prior to the original tracking, learning is carried out in offline-learning way. In order to learn a similarity function among tracklets and detections, ground truth trajectories supervision is used by [5] for association of data. In order that, offline learning is static and it cannot consider the target history as well as dynamic status in association of data that is essential to solve ambiguities. It is mainly required to reallocate the occluded or missed objects while it appears again. While tracking, learning is conducted by online-learning. In order to the tracking outcomes, a general method is to build on negative as well as positive training instances and to

similarity function training for associating the data. Depending on the status and target history, online-learning is capable to use the features. But, there exist no real truth annotations for supervision. From inaccurate training instances, the techniques have a probability to learn when there exist mistakes in tracking outcomes. It might be collected and gives a tracking drift. In Markov Decision Processes (MDPs), we form online multi-object tracking problem, wherever the object lifetime is designed with MDP, and for multi-object tracking, the MDPs are grouped.

Recent MOT studies aims at the principal of tracking-by-detection using the major differences in regions. In a graph-based representation, most of the batch techniques [6] are presented. The online techniques resolve the data association issue either determinatively or probabilistically [7]. In any data association method, a main element is similarity function among objects. Online and batch techniques [8] had discovered the learning idea towards tracking, wherever the target is to similarity function learning for association of data out of the training data. The major work contribution is a new reinforcement learning technique for association of data within online MOT. The standard trackers aim at how to learn the strong target online appearance model and employ it for tracking. As they are not capable to manage exiting/entering the scene object, it is difficult to use the trackers towards MOT. Towards various tasks of computer vision, the processes of Markov decision [9] had been used like human activity forecasting, human-machine collaboration and feature selection for recognition. For dynamic framework, the MDP is appropriate wherever an agent requires carrying out particular jobs through executing activities and creating decisions consequently.

From the availability of MOT applications on different domains, in this study, the problem of anomaly detection in pedestrian walkways is considered. Recently, deep learning based learning [10] based anomaly detection models are presented. At the earlier days, convolution neural network (CNN) [11] is applied for classifying the occurrence of objects in region as an anomaly or not. But, it faces the problem of different spatial positions of the objects present in the image. Therefore, it is required to select the region count and it results in high computational complexity. To reduce the complexity level, region based CNN (R-CNN) [12], YOLO, etc are evolved for identifying the existence of objects in a rapid way. For resolving the difficulty of selecting massive regions, a novel way of using selective search technique is employed to filter out only 2000 regions called region proposals. It leads to slight reduction in computation complexity of CNN. Simultaneously, the time required to train the CNN leads to improve the need of classifying 2000 region new models only for specific image. Also, it is not possible to install in real-time applications since it takes longer duration for individual test image.

The upcoming portions are structured here. Section 2 elaborates the IRS-CNN model and Section 3 describes the experimental investigation. Then, Section 4 draws the conclusion.

II. Improved Region Based Scalable Convolution Neural Network (IRS-CNN)

The traditional MOT formulation is employed in this study that generates interested objects as outliers. At the presence of standard conditions, a statistic approach $p_x(x)$ is defined to distribute the measurement X in a normal constraint. Here, in anomaly detection, the abnormal objects are represented as measurements which has the probability lesser than a threshold value. It is equivalent to a statistical test of hypothesis:

$$\mathcal{H}_0: \text{x is attained out of } p_X(x) \quad (1)$$
$$\mathcal{H}_1: \text{x is taken from an uninformative distribution } p_X(x) \propto 1$$

When $p_X(x) < v$, where v denotes a normalization constant to avoid null hypothesis $\mathcal{H}_0$.

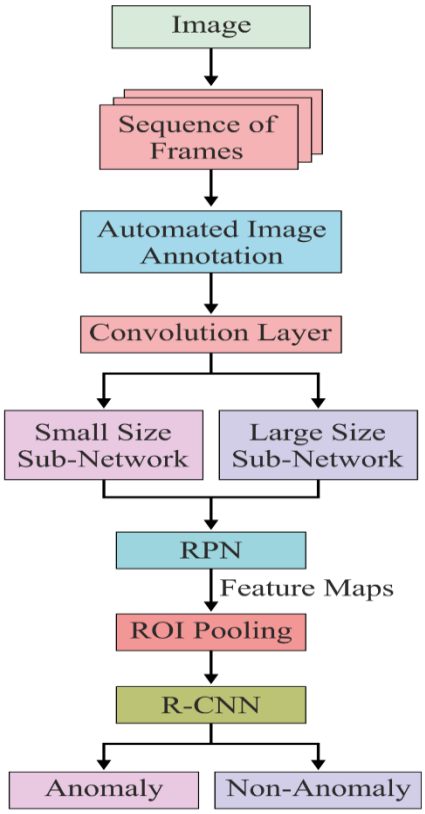

Fig. 2. The overall process of IRS-CNN model

The entire procedure of the presented IRS-CNN technique is provided in Fig. 2. The IRS-CNN technique comprises of three major stages: AIA model for annotating images, RPN with scalable features and Fast R-CNN. The first stage the generation of image annotations, second stage involves a deep fully CNN that creates the regions and the third stage utilizes the produced regions for detecting anomaly. The RPN along with scalable features commands the Fast R-CNN model for searching the object. For resolving the challenges exists in earlier models, the presented IRS-CNN model holds of significant size and smaller size sub-

networks that detect the anomaly in various sizes. The classifier outcomes are applied to the RPN for generating region proposals. The scaling aware integration with AIA creates the presented IRS-CNN technique to efficiently detect the abnormalities of different size and distributes the convolution filters from existing layers.

The IRS-CNN technique depends upon a popular Faster R-CNN detecting technique due to the fact that the improved detection results are attained with less computational complexity. The IRS-CNN model acquires the whole image as well as values of object proposals in the form of input to predict the abnormal issue present in a input image. At the initial level, IRS-CNN segmentation task will be carried out where the input videos undergo segmentation for the identification of anomaly regions. At this point, the IRS-CNN features are determined to extract the correlation. Next, the AIA process will be carried out. Then, the identification of anomalies will be carried out. The actual input for the identification of anomalies is the sequence of frames in the videos in addition to detected regions. The outcome can be the anomaly identified frames with respective label. To carry out this process, the features of IRS-CNN will be filtered and observed regions will be mapped. Once the observed regions are determined in the frame, the respective labels will be assigned with its detection rate.

A. AIA based on CNN_WARP

To develop an AIA model for annotating images in a faster automated way, CNN with Weighted Approximate Ranking (CNN_WARP) is employed. The robust visual features seem to be the basic factor to annotate mages. The conventional handcrafted features are stochastic and not satisfactory. Based on the successful performance of CNN in the domain of computer vision, CNN is employed for generating stronger visual characteristics for AIA. CNN+WARP model [14] make use of ranking process for training deep CNN for AIA. Here, a set of 5 five convolution layers and 3 widely connected layers from CNN model is applied. To define a collection of images x, the convolution network is represented as f(.). The outcome off(.) denotes a scoring function of data point x which has been comprised with a vector of activation. It is also considered that n images and c tags are applied to train it. Using Eq. (2), the WARP loss function will be minimized:

$$K = \sum_{i=1}^{r} \sum_{j=1}^{c+} \sum_{k=1}^{c-} L(r_j) \max(0, 1 - f_j(x_i) + f_k(x_i) \quad (2)$$

where L(.) implies the weighting function to represent diverse ranks, and r_j represents the rank for j^{th} class for image i. Hence, weighting function L(.) can be determined as:

$$L(\cdot) = \sum_{j=1}^{r} a_j \quad (3)$$

Where a_j is represented by 1/j and the weights L(.) manage the top-k of the optimization. Additionally, the rank r_j is determined using Eq. (4) for classes $c$and sampling trials s.

$$r_j = \left| \frac{c-1}{s} \right| \quad (4)$$

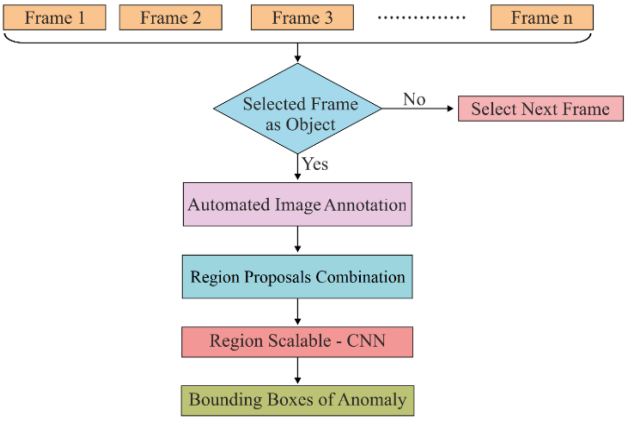

Fig. 3. Identification of bounding boxes

B. *Region Proposal Networks*

An RPN obtains an input image with diverse sizes where the attained results are in the form of collective rectangular object proposals with exclusive objectless value. Since the main theme of this model is based on resource sharing with Fast R-CNN, it has been declared that each net shares similar set of convolution layers. In order to develop the region proposals, tiny network gets the input as $n \times n$ spatial window. All sliding windows are mapped to a low-dimensional feature which provides a 2 fully connected layers such as, box-regression layer (reg) as well as box-classification layer (cls) as illustrated in Fig. 3.

Anchors

As a summary, the RPN carry out a grading procedure for giving ranks to every region box (called as anchors) and represents the possible one which hold the objects. The anchor acts as a significant part in the implementation of Faster R-CNN which is considered as a box with a 9 anchors exists in the image positions. Fig. 4 depicts the deployment of 9 anchors at (320, 320) location of an image with the size of (600, 800).

Loss Function

For training RPN, it is essential to assign a binary class label either it is normal or abnormal for all anchors. It has been monitored that unique ground truth box positive labels are allocated to several anchors. A negative label is scheduled for minimum IoU ratio (under 0.3) for each ground-truth box. Therefore, anchors does not belongs to positive or negative anchors.

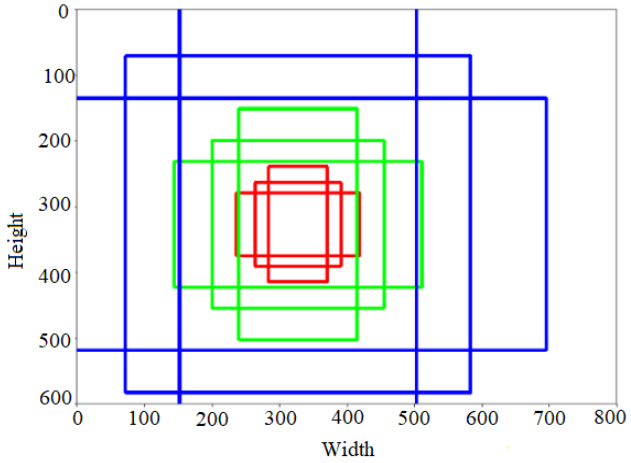

Fig. 4. Anchor

The Faster R-CNN tends to decrease the objective function applied by a multiple loss in Fast R-CNN. Hence, the loss function of the image has been expressed in Eq. (5):

$$L(\{p_j\},\{t_j\}) = \frac{1}{N_{cls}}\sum_j L_{cls}(p_j, p_j^*) + \lambda\frac{1}{N_{reg}}\sum_j p_j^* L_{reg}(t_j, t_j^*) \quad (5)$$

where j is anchor index in a minibatch and p_i is the determined probability of anchor j has been declared as an object. The value of ground-truth label p_j^* can holds the values of 0 (negative anchor) and 1(positive anchor). t_j signifies a vector of 4 parametric coordinate points, t_j^* is a ground-truth box has been correlated to a positive anchor. The classification loss L_{cls} is log over 2 classes. The regression loss is indicated as $L_{reg}(t_j, t_j^*) = R(t_j - t_j^*)$, where R implies the robust loss function (smooth L_1). The element $p_j^* L_{reg}$ represents the regression loss that lies in the inactive state ($p_j^* = 1$) and active state ($p_j^* = 0$). The outcome from the cls and reg layers holds $\{p_j\} and \{t_j\}$. The parameters involved in bounding box regression are represented in Eq. (6):

$$t_x = (x - x_l)/w_l, \quad t_y = (y - y_l)/h_l$$
$$t_w = log\,(w)/w_l, \quad t_h = log(h)/h_l \quad (6)$$
$$t_x^* = (x^* - x_l)/w_l, \quad t_x^* = (y^* - y_l)/h_l$$
$$t_w^* = log(w^*)/w_l, \quad t_h^* = log\,(h)/h_l$$

where x, y, w and h are the box's center coordinates, width and height. The variables$x$, $x_l$andx^* represents the predicted box, anchor box and ground truth box.

C. *ROI Pooling*

The simulation outcome attained from RPN provides various size of proposed regions. These diverse sized regions shows varied sizes CNN feature map. It is very difficult to create a productive model to operate the features with different sizes. This technique is capable of reducing the complexity in feature map. In contrast to Max-Pooling, ROI pooling classifies the feature map to specific number of regions. Then, Max-Pooling has been applied for all regions. Thus, the result obtained from ROI Pooling is often a k regardless of input size.

D. *Sharing Convolutional Features for Region Proposal and Object Detection*

For detecting the anomalies present along the pedestrian walkways, Fast R-CNN approach is applied. The process of making a method to learn the convolution layers that is shared among RPN and Fast R-CNN are discussed. In addition, the convolution layers can be altered in different forms. It is essential to create a technique with a distribution of convolution layers from 2 networks instead of processing a learning process which has to be conducted at 2 various systems. It has been assumed that presenting an individual network with RPN and Fast R-CNN is very tedious and optimizing the back propagation (BP) model. A 4-step training approach is employed to learn the shared features by the optimization process. At the beginning, RPN undergo training process. At the next stage, separate detection network undergo training by the use of Fast R-CNN making use of proposals are produced from RPN at initial phase. Here, RPN as well as Fast R-CNN could not distribute the convolution layers. A detecting network were utilized in the upcoming stage to initiate a training process of RPN. At the same time, a shared convolution layers are predetermined which has the exclusive layers from RPN to be tuned in a smooth fashion.

III. Conclusion

The rise of big data in the field of video surveillance has created a need to develop an efficient identification model to track the moving objects. Since the computation complexity and the time taken for image annotation is found to be high, this paper has introduced an IRS-CNN based MOT model. The presented IRS-CNN model enhances the existing RS-CNN by incorporating AIA tool for increasing the detection rate as well as reducing the computation time. To develop an AIA model for annotating images in a faster automated way, CNN_WARP model is employed. The interesting feature of AIA tool helps to rapidly annotate the training images in an automatic way. This automated annotation tool requires manual annotation of region of interest in one frame and automatically annotate the similar targets in the rest of the frames rather than the requirement of manually annotating the targets in all the frames in the applied video. This interesting feature of the proposed model helps to reduce the labor and annotation time. Besides, the automatic annotation process enables the proper identification of objects in the training process and thereby detection rate can be significantly increased. In this study, a testbed of Test004 video sequence from UCSD Anomaly Detection Dataset is employed. An broad experimental result verified the optimal behavior of IRS-CNN model against a set of applied test images over the compared methods.

REFERENCES

[1] Gandomi A, Haider M (2015) Beyond the hype: big data concepts, methods, and analytics. Int J of Info Man 35:137–144.

[2] MDP Tracking. http://cvgl.stanford.edu/projects/ MDP_tracking.

[3] Multiple object tracking benchmark. http://motchallenge. net.

[4] B. Babenko, M.-H. Yang, and S. Belongie. Robust object tracking with online multiple instance learning. TPAMI, 33(8):1619–1632, 2011.

[5] S. Kim, S. Kwak, J. Feyereisl, and B. Han. Online multi-target tracking by large margin structured learning. In ACCV, pages 98–111. 2012

[6] Y. Li, C. Huang, and R. Nevatia. Learning to associate: Hybridboosted multi-target tracker for crowded scene. In CVPR, pages 2953–2960, 2009

[7] Z. Khan, T. Balch, and F. Dellaert. Mcmc-based particle filtering for tracking a variable number of interacting targets. TPAMI, 27(11):1805–1819, 2005.

[8] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. TPAMI, 34(7):1409–1422, 2012.

[9] R. Bellman. A markovian decision process. Journal of Mathematics and Mechanics, 6(5):679–684, 1957

[10] LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. nature, 521(7553), p.436.

[11] Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural networks, 61, pp.85-117.

[12] Girshick, R., 2015. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

[13] Ren, S., He, K., Girshick, R. and Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems (pp. 91-99).

[14] L. Wu, R. Jin, A. K. Jain, Tag completion for image retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (3) (2013) 716.

[15] UCSD Anomaly Detection Dataset, http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm

[16] Murugan, B.S., Elhoseny, M., Shankar, K. and Uthayakumar, J., 2019. Region-based scalable smart system for anomaly detection in pedestrian walkways. Computers & Electrical Engineering, 75, pp.146-160.

[17] Mahadevan, V., Li, W., Bhalodia, V. and Vasconcelos, N., 2010, June. Anomaly detection in crowded scenes. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on (pp. 1975-1981). IEEE.

[18] Kim, J., and K. Grauman., 2009. Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates. In CVPR, pages 2921–2928.

[19] Mehran, R., Oyama, A., and Shah, M., 2009. Abnormal crowd behavior detection using social force model. In CVPR, pp. 935–942.