MULTIPLE OBJECT DETECTION IN SURVEILLANCE VIDEO WITH DOMAIN ADAPTIVE INCREMENTAL FAST RCNN ALGORITHM

Fancy Joy

Dr. V. Vijayakumar

1,2 Department of Computer Science, Sri Ramakrishna college of Arts and Science Coimbatore, India.

Abstract

Object detection is one of the key steps in wide variety of surveillance-based applications. This paper introduces an incremental class based Fast RCNN multiple object detection method that can be apply in various domain of surveillance video analytics. The two major components of proposed method are class incremental learning component and domain adaptation component. Class incremental learning helps for incremental adding of object classes with existing object classes for Fast RCNN. Domain adaptation component part improves the detection even the object presents in cross domain. Here basic object classes can be trained initially and then can add new object class incrementally according to new class requirement or according to domain. Benchmark datasets COCO, PETS S2L1, CD2014, Visor dataset were used for training and testing. Experimental results were showed that the proposed approach performs better in video datasets that contains challenges such as illumination variation, presence of shadow, partial occlusion and dynamic background.

Keywords: Multiple object detection; incremental learning; domain adaptive; Fast RCNN; mean Average Precision (mAP).

1. Introduction

Detecting objects in surveillance video plays a vital role in surveillance video analytics and have many applications in various domains such as object tracking, suspicious event detection, navigation and security etc. [1]. Object detection in surveillance video identifies the objects in each frame and locate the object instances of appropriate categories in an image. Detection of multiple objects becomes complex due to various challenges such as varying illumination, weather condition, presence of shadow, fog etc. [2].

Some of the object detection methods are background subtraction, frame differencing, optical flow, Fast RCNN, Faster RCNN etc. Each has its own advantages and limitations. Recent trends in this area are based on deep learning and achieved better results on benchmark datasets such as Pascal VOC and COCO [3][4][5]. Due to the evolving nature of datasets with new classes and samples arises the necessity of incremental learning method. Incremental learning method helps to add new classes with existing object classes incrementally without training from scratch. The proposed method can detect object and predict corresponding object class. It gives efficient detection and classification even in dynamic background, presence of shadow, occlusion and illumination variation.

Object detection algorithms were mainly categorized in to background subtraction, frame difference, optical flow and convolutional neural network (CNN) based. Majority of the recent works in object detection were in deep learning [3][4][5]. Some of the deep learning-based methods are RCNN, Fast RCNN, Faster RCNN, Mask RCNN etc.

Region based Convolutional Neural Network (RCNN) [6] achieves good object detection accuracy using deep convolutional neural network to classify objects. But RCNN consumes more time for training and detection. The author R Girshik[4] proposed fast RCNN in order to overcome the limitation of RCNN. He used selective search method to obtain bounding box and CNN for feature extractions. Hence improved speed of training and testing and detection accuracy. Further improvements require new techniques that allow both dense and sparse proposals detection. R Girshik et al. [5] again introduced the concept of Region Proposal network to generate the bounding box proposals in place of selective search and built an end to end detection framework, Faster RCNN. Shelmokov et al. [3] developed a Fast RCNN based class incremental model for object detection. The bounding boxes were...
produced by Edgeboxes. The method can incrementally add new classes with preserving performance of old classes. The method can further enhance by adapting the method to learned proposals. Chen et al. [7] built end to end model for cross domain object detection based on faster RCNN. It used image and instance level domain discrepancy in detection. They were validated in various domain shift scenarios and proved the efficiency in cross domain detection.

M H Nguyen et al. [8] discussed compressed domain approach for detection of object. The approach can detect the interested regions fast even with uniform color objects. Extensions required for accurate object detection using automatic threshold. N.S Ghedia et al. modified Gaussian mixture model for indoor and outdoor surveillance video datasets [9][10]. The author made intrinsic and extrinsic improvements like adaptive threshold and other parameters. It detects objects in various circumstances. Even though, it requires extensions for detection of objects in presence of complete shadow and full occlusion. Huang et al. [11] demonstrated a method based on optical flow for object detection in unconstrained scenes. The approach adapts to itself to various scenarios and it outperforms than the state of art method such as SA (Stochastic Approximation), SCBU (Scene Conditional Background Update). Ray and Chakraborty [12] proposed object detection method suitable for dynamic environment. Pseudo motion in background was estimated by phase correlation of adjacent frames based on Fourier shift theorem and fixed the background. Using the result, the algorithm efficiently separates foreground from background and this Foreground is termed as objects. But it is difficult to handle the objects in very crowded scenario and extreme illumination variation.

The authors Mahalingam and Subramoniam[13] presented robust moving object detection method consists of detection, tracking and evaluation phase. They used Mixture of adaptive Gaussian for foreground segmentation and decision tree-based classifier for feature extraction and classification respectively. Detecting objects in crowded scene, full occlusion and extreme illumination variation might be improved. Goyal and Singhai [14] used Gaussian mixture model with foreground matching. This algorithm detects slow moving objects effectively. They adopted adaptive learning rate to deal with different rate of change of background. Improvements will be needed to deal with detection in presence of shadow. Detection of objects based on color histograms and local binary patterns were introduced by Kwon lee et al. [15]. Method generated the color histogram for image and reduces its bins. Then its small number of bins again matched with LBP histograms for object detection. Results were shows that method effective and fast.

Some of the object detection approaches based on domain adaptation [16][17][18][19] supports object detection in cross domain. Saito et al. [16] proposed adaptive object detection based on strong local alignment weak global distribution alignment. Y. Ganin et al. [17] discussed unsupervised object detection method based on deep architectures. The method achieved adaptation effect based on big domain shifts. X. Zhu et al. [18] developed region level adaptive framework for object detection. The method discriminates the region and adjusted region level alignment. The achieved efficiency and good scalability in object detection on various domain shift scenarios. Object detection by learning paradigms were introduced by T. Kim et al. [19] in 2019. Domain diversification discriminate objects using deep feature extraction. Multi domain learning paradigms for adversarial learning for cross domain object detection.

Deep learning methods for object detection mainly focus on the features of objects and detect object according it. Most of the other methods based on background subtraction concept analyses the background and extract the foreground consider as object. An incremental detector is necessary for surveillance video object detection which helps to add more classes or customize classes flexible to target domain. An image level domain adaptive method was proposed here which is purely based on incremental fast RCNN method. The fast RCNN with incremental learning method helps to train new object classes without difficulty. The proposed incremental fast RCNN detects the moving objects even in cross domain to an extent. Experimental results on the surveillance benchmark datasets such as PETS, VISOR and Change Detection 2014 clearly demonstrates that the proposed approach shows better performance in object detection of newly trained classes as well as previous classes even in domain shift scenarios. The main objective of the proposed method is to detect the multiple objects in surveillance video, which is significant in the area of surveillance video analytics. The contributions of the papers are follows

- Generic cross domain multiple object detection method suitable for various surveillance video applications.
- End to end class incremental object detection and classification without catastrophic forgetting
- Better object detection performance even in different challenging scenarios such as illumination variation, presence of shadow, fog and in-consistent object motion etc.
- Class incremental method ensures the scalability. New classes can be trained without from scratch if application needs.

The organization of the paper as follows: section 2 and 3 discusses class incremental domain adaptive fast RCNN, experimental results respectively. Finally, section 4 summarizes the conclusion and future work.
2. Class Incremental Domain Adaptive Fast RCNN

2.1. Fast RCNN

Fast RCNN [4] is the base network for proposed approach developed by R. Girshik in 2015. Method used convolutional neural network for feature extraction, softmax SVM (support vector machine) for classification and outputs object detection with bounding boxes and classes. Fig 1 represents the structure of Fast RCNN [4].

![Fast RCNN Diagram](image1)

Here Fast RCNN is the base network for proposed approach. The network process whole image through a sequence of residual networks and generate corresponding feature maps. Simultaneously 2000 pre-computed object proposals were also generated by using MCG proposals [20]. Resnet 101 is a deep residual network use as feature extractor CNN. The combinations of both feature maps and object proposals uses for ROI projection and perform ROI pooling. These passes through fully connected convolution layer and outputs object classification and object localization.

2.2. Class incremental domain adaptive fast RCNN

The proposed method implements class incremental multiple object detection for surveillance video. Incremental learning with fast RCNN helps easy and fast training of object with classes which are required in surveillance video. For implementation method requires two fast RCNN networks: One for base class training and other network for addition of new classes, when requirement arises. Image level domain adaptation also uses here to improve detection efficiency. The overall architecture of proposed model is depicted in Fig 2. The method is purely based on Fast RCNN. It includes an original detector Fast RCNN (Network A) used to train the old classes and network B for new class adaptation achieved by the incremental learning i.e. training without from scratch.

![Incremental Fast RCNN with domain adaptation](image2)
The model chose Fast R-CNN with deeper ResNet-101 feature extractor component, which is faster and more accurate than VGG-16. The input image and about 2000 pre-computed object proposals represented as bounding boxes were given as input to Fast RCNN network. It passes through a series of residual blocks and then ROI pooling layer performs max pooling over the proposals to generate 7×7 feature maps. The method used two fully connected layers such as soft max layer for classification and regression layer for bounding box localization. A per category Non-Maxima Suppression (NMS) is applied here to get accurate detection result.

The training of proposed method can be achieved by dual network learning method. The entire training framework is divided into two set of classes A and B. Class A and B includes various set of classes. Initially train the model first set of classes to network A (N_A). Secondly aims to add new set of classes using Network B (N_B) without training from scratch. Concatenate the Network N_B with Network N_A and the retraining process (i.e. adding new classes) is repeated whenever new class requirement arises. For adding new class C to existing network train class C separately (N_C) and integrate with previously trained classes N_AUN_B.

Lower part of the architecture includes domain adaptation component. Image level domain classifier added after the last convolution layer of fast RCNN. Gradient Reverse Layer (GRL) [21] use here to train domain adaptation components and it automatically reverse the gradient during propagation. An image level domain shift component incorporated with the incremental fast RCNN method to minimize the H -divergence between two domains. H -divergence is the divergence measure between two different distributions. To reduce the domain mismatch problem the method employed a patch-based domain classifier.

The overall loss of incremental Fast RCNN includes two type of losses Classification loss \( L_{cls} \) and regression loss \( L_{loc} \). Classification loss means how accurately predicts class of objects and the regression loss denotes how correctly refine the bounding box location.

\[
L_{CIFN} = L_{cls} + L_{loc} \tag{1}
\]

2.3. Image level domain adaptation

Consider ‘I’ is the image representation i.e. output of base convolutional layers of feature map. Input image or frame represented as convolutional feature map shared by input convolutional layers. Based on the feature map ROI classifier predicts the category label from ROI pooling. Image based domain adaptation used in this method to handle the domain inconsistency. Domain adaptation in image level improves the detection in various domain shifts (e.g., Image scale, image style, illumination, etc.).

Domain adaptation component part takes the feature map and predicts the domain label for each image patch. Based on feature map and image patch, the domain classifier trained. The process repeats for each image patch. The main advantage of this patch-based classifier is that it can deal with the global transformation of the image (image scale, illumination variation etc.). Gradient decent method used here for train the domain adaptation. In implementation a GRL (Gradient Reversal Layer) used between feature extractor and domain classifier to ensure feature distribution over two domains are similar.

Image level domain classifier linked with the ROI wise features of Fast RCNN. Consistency loss is used here to make proposed Fast RCNN approach to be domain invariant. The overall training loss of the proposed network can be calculated as

\[
L = L_{CIFN} + \lambda(L_{img}) \tag{2}
\]

Where, \( \lambda \) indicates the trade-off parameter to balance the Fast RCNN loss and domain adaptation component. \( L_{CIFN} \) loss of Fast RCNN from (1).

Image level domain adaptation indicates the ability of proposed object detector that can detects the same object even though it is in different domain, then it can also able to detect the object I if it is present in other domain PR(I) is referred as image level domain adaptation. Here approach uses the domain classifier and adversarial training for domain invariant robust feature training.

In probabilistic perspective of image level domain adaptation can be represented as posterior P using Bayes’s theorem. Consider two domains R, S and I is the representation of image in both R and S. B indicates bounding box and C for object classes.

Then method trains object I in domain S (source domain). Conditional probability \( P_S(C,B,I) \) can be represented as

\[
P_S = P(C,B,I) = P(C,B/I) P(I) \tag{3}
\]
Image representation in target domain conditional probability $P_T(C, B, I)$

$$P_T = P(C, B, I) = P(C, B|I)P(I) \quad (4)$$

Here the domain is different but conditional probability $P(C, B|I)$ is same for both domains. Hence change the shift mainly occur in image distribution $P(I)$. The approach can detect the object regardless the domain which the object belongs (i.e. the result must be same in cross domain). To handle this scenario approach, enforce that the $P(I)$ from two domains as same i.e. $P_S(I) = P_T(I)$.

3. Experimental Results and Discussion

3.1. Experimental setup and dataset

The main objective of the proposed method is to detect and classify multiple objects correctly. Dataset used for training and testing are COCO and various surveillance video datasets PETS S2L1[22], visor[23] and Change Detection (CD 2014)[24]. For train, validation and test of the method uses specific classes of COCO with ground truth of PETS 2009, VISOR dataset. The method trained and tested using first 8 classes of COCO dataset. Classes were divided into two categories. Category A includes car, person, bicycle, truck and category B includes airplane, motor bike, bus, train. The network A trained using category A set of classes with Fast RCNN method. Network A trained with 40K epochs with training category A classes with learning rate 0.001, momentum 0.9 and elu activation function. In the second stage of training category B classes were trained with parameters 20K epochs, learning rate 0.0001 and momentum 0.9. $\lambda=0.1$ for domain adaptation training and testing.

3.2. Evaluation metrics

For all experiments the method used 0.5 as threshold for mean average precision (mAP) evaluation. Following are the basic notations for classification.

TP$_C$-Correct prediction of class
FP$_C$-Wrong prediction of class
FN$_C$-not detected class even though exist class in ground truth

Performance metrics for object detection and classification are follows.

- Average Precision for a class $C$ is follows
  $$AP_C = \frac{TP_C}{TP_C + FP_C} \quad (5)$$

- Mean Average Precision refers to the average precision of all classes.
  $$mAP = \frac{1}{K} \sum_{i=1}^{N} AP_i \quad (6)$$

Basic terminologies for object detection are:
TP- Correct detection
FP- Wrong detection
FN- Not detected existing object in ground truth

The performance metrics used for object detection in surveillance video are Precision, Recall and F-Measure. Precision ($P$) is the percentage of correct positive detection to all detection and recall ($R$) is the percentage of correct positive detection to the ground truth elements.

$$Precision(P) = \frac{TP}{TP + FP} \quad (7)$$

$$Recall(R) = \frac{TP}{TP + FN} \quad (8)$$

$$F - Measure = \frac{2(Precision \times Recall)}{Precision + Recall} \quad (9)$$

3.3. Experimental results

In this section the proposed approach evaluated using above mentioned performance metrics equations.

3.3.1. Object detection and classification
Fig 3 shows some example frames clearly demonstrate that the proposed approach can efficiently detect classes in both categories A and B. Table 2 indicates the object detection and recognition results of coco dataset. The method other than proposed approach used selective search algorithm for proposal generation and VGG16 for feature extraction. Our method used pre-computed MCG proposal to get the object proposal and Resnet-101 for feature extraction. Hence proposed approach shows improved mAP in table 1.

<table>
<thead>
<tr>
<th>Method proposals</th>
<th>COCO test (mAP@.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fast RCNN SS,2000</td>
<td>35.9</td>
</tr>
<tr>
<td>Fast RCNN(faster rcnn paper) SS,2000</td>
<td>39.3</td>
</tr>
<tr>
<td>Incremental Fast R-CNN+DA (Proposed) MCG,2000</td>
<td>46.18</td>
</tr>
</tbody>
</table>

Table 1. Object detection results on COCO dataset

In table 2 the method compares result of incremental fast RCNN without distillation [6] and proposed method i.e. incremental fast RCNN with Domain adaptation without distillation. The result shows that the average precision of class of proposed approach better than other.

<table>
<thead>
<tr>
<th>Method</th>
<th>Aero</th>
<th>mbike</th>
<th>Bus</th>
<th>Car</th>
<th>truck</th>
<th>person</th>
<th>Train</th>
<th>bicycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incremental Fast RCNN without Distillation</td>
<td>35.9</td>
<td>36.1</td>
<td>26.4</td>
<td>36.2</td>
<td>_</td>
<td>9.1</td>
<td>28</td>
<td>_</td>
</tr>
<tr>
<td>Incremental Fast RCNN without Distillation +DA Proposed</td>
<td>86.6</td>
<td>48.16</td>
<td>62.2</td>
<td>58.9</td>
<td>63.2</td>
<td>69.93</td>
<td>33.7</td>
<td>29.23</td>
</tr>
</tbody>
</table>

Table 2. COCO+PETS2009+Visor test per class average precision when 8 classes added

3.3.2. Object detection in surveillance video datasets

The proposed method tested the detection result in various surveillance video datasets. The chosen videos include various challenges such as illumination variation, presence of shadow, occlusion and dynamic background etc. Approach compares with benchmark Gaussian Mixture Model implemented by Stauffer and Grimson [25] for performance evaluation. Figure 4 depicts the precision and recall graph for PETS S2L1 video dataset and it concludes that proposed approach shows better precision and recall value for most of the video frames of PETS S2L1. The approach gives less false positives than GMM.
F-Measure values for various challenging video sequences were illustrated in Fig 5. F-Measure result of proposed approach proves that the proposed approach is efficient than GMM in all challenging video sequences. Below graph displays the F-Measure results of video sequences from PETS, VISOR, CD 2014 datasets.

Figure 6 demonstrates one of the main advantage incremental fast RCNN with domain adaptation. The image domain adaptation part helps the method to detect the object even though it is in cross domain. The approach can detect the same object regardless its background.
4. Conclusions
This paper presented a novel object detection method can be used in surveillance video analytics. The proposed approach includes class incremental method, which helps the algorithm to cope up with new object requirements. Domain adaptive nature of proposed method makes it adaptive to various weather conditions to an extent. Initially basic set of classes can be trained and deployed. Then the approach can extend with more classes according to the application domain without training from scratch. The approach gave efficient detection in various video datasets. It reduced the number of false positives. But the limitation is that the method shows poor detection in far-away objects and full occlusion. Further enhancements are required to speed up the algorithm by replacing Fast RCNN with Faster RCNN method.

Acknowledgments
We thank Sri Ramakrishna College of arts and science for giving support for doing this research work.

References
[23] http://imagelab.ing.unimore.it/visor/
Authors Profile

**Fancy Joy**, Research Scholar, Sri Ramakrishna College of arts and Science. Published 6 papers in International/ National Conferences and Journals. Research interest includes Machine learning and deep learning.

**Dr V.Vijayakumar**, Professor, Sri Ramakrishna College of Arts and Science, Coimbatore. He has twenty years of experience in Teaching, Research & Development. He awarded Research degree in the area Video Data mining. He has recognized as Research supervisor in Bharathiar University and guided 8 M.Phil scholars. Currently, he is guiding three Ph.D Scholars and one M.Phil Scholar. He completed AICTE Research Promotion Scheme grant to the tune of Rs.10.5 lakh and currently received Major Research Grant Rs. 10.5 Lakh form ICSSR-IMPRESS. He was appreciated by the Analog and Digital Systems. He was honoured with Best Academic Officer Award by EMC Corporation Bangalore. He has published thirty papers in International/ National Conferences / Journals including Springer and IEEE. He has delivered lectures in Big Data Analytics (Hadoop and Ecosystems), R-Programming, No-SQL (MongoDB), Internet of Things and SMACT. He is serving as Subject Matter Expert and Adjunct Faculty in Texila American University, Guyana, South America. He is a reviewer of various International Conferences and journals. He is serving as a member of various bodies. His research interest includes Machine Learning, Computer Vision, Data Science & Big Data Analytics and Internet of Things. He is a member in ISTE, IAENG, IACSIT and CSTA professional societies.