Improving Real-time Pedestrian Detection using Adaptive Confidence Thresholding and Inter-Frame Correlation

1st Mufleh Al-Shatnawi
Electrical Engineering and Computer Science
York University
Toronto, Canada
mufleh@eecs.yorku.ca

2nd Vida Movahedi
Electrical Engineering and Computer Science
York University
Toronto, Canada
vida@eecs.yorku.ca

3rd Amir Asif
Electrical and Computer Engineering
Concordia University
Montreal, Canada
amir.asif@concordia.ca

4th Aijun An
Electrical Engineering and Computer Science
York University
Toronto, Canada
aan@cse.yorku.ca

***Abstract*—The pedestrian detection algorithms form a key component in the multiple pedestrian tracking (MPT) systems. Despite efforts to detect a pedestrian accurately, it is still a challenging task. We propose a novel and efficient online method to improve the performance of the multiple person/pedestrian detector by introducing novel post-processing steps. These steps use an adaptive approach to determine both area and confidence score constraints for the output of any given multiple pedestrian detector. In this paper, we focus on pedestrian detection in video surveillance applications that require an automated, accurate and precise pedestrian detection algorithm. We demonstrate that the new steps make the multiple pedestrian detector more accurate, precise and tolerant to false positive detections. This is illustrated by evaluating the performance of the proposed method in test video sequences taken from the Pedestrian Detection Challenge, Multiple Object Tracking Benchmark (MOT Challenge 2017).**

***Index Terms*—Multiple object detection, Multiple pedestrian detection, Convolutional neural network, Video surveillance.**

I. Introduction

Pedestrian detection algorithms have received considerable attention recently. The pedestrian detection results are used in a wide range of applications in computer vision, such as video surveillance, traffic safety, vehicle navigation and sports analysis. In these applications, the pedestrian detection algorithm should be accurate and precise. Despite efforts to generate accurate and reliable pedestrian detections, it is still a challenging task for researchers to develop a perfect Multiple Pedestrian Detector (MPD).

Normally, MPDs produce both a bounding box and confidence score for each detected pedestrian in a given video frame. The confidence score represents the confidence level of the detector in affirming that the object enclosed by the bounding box is a person/pedestrian. The traditional approach for pedestrian detection is based on background-subtraction [1]- [4]. In this approach, pedestrians are detected in every frame by segmenting the moving objects out of the background, while taking into account pixel-wise time consistency. However, the background-subtraction methods are unreliable and error-prone in noisy video sequences. For instance, the background-subtraction methods detect all moving objects in the scene even these that are not pedestrians [1]- [4].

In recent years, multiple pedestrian detection methods have been developed either by using a deep Convolutional Neural Network (CNN), or by building a specific pedestrian detector added to these networks [5]- [10]. These CNN pedestrian detection methods are able to learn discriminative features directly from raw pixels of an image, and they are producing a confidence score between zero and one for the detected pedestrians. Hence, these methods have notable performance gains over the background-subtraction methods, and they normally provide a high detection accuracy. Moreover, the CNN pedestrian detection methods are generally robust to changing background and to camera motion as compared to background-subtraction methods.

In [7], a pedestrian detector is proposed by using the Faster Region Convolutional Neural Network (Faster-RCNN). The Faster-RCNN can be represented as an end-to-end framework that consists of two sub-CNN networks. The first network extracts features and proposes regions for the second network which in turns classifies the object in the proposed relevant regions. The Faster-RCNN parameters are shared between these two networks and constitute an efficient framework for object detection in general. Furthermore, the Faster R-CNN can be viewed as a CNN based MPD without using any hand-crafted features. The performance of this MPD based approach, here referred to as FRCNN, is evaluated using the pedestrian detection challenge in multiple object tracking benchmark (MOT Challenge 2017) [11]. The confidence scores of the reported pedestrian detections were between 0.05 and 1.0. In [8], another MPD based approach is developed by using a combination of an additional convolutional neural network and the Faster R-CNN [7]. The additional network is used to calculate the appearance descriptor value for each detected bounding box. The calculated value is then used to determine the data association metric for later stages. The performance of this MPD, here referred to as KDNT, is evaluated using

the pedestrian detection challenge in multiple object tracking benchmark (MOT Challenge 2017). The confidence scores of the reported pedestrian detections were between 0.0990 and 0.9998. The KDNT detector was ranked as number one in the MOT Challenge 2017 as of the writing of this paper (February, 2018).

In benchmark datasets, the pedestrian detector performance is evaluated by comparing the reported bounding boxes with the ground truth (GT) bounding boxes for each video frame. The performance on each individual frame is then averaged over all the frames to obtain the final performance score of the pedestrian detector [12]. Three main performance metrics are calculated for each video frame. These metrics are: (i) True positive (TP) representing the number of pedestrians/bounding boxes that are detected and comply with the ground truth (GT) (also called correct detection), (ii) False positive (FP) representing the number of pedestrians that are detected but not present in the GT (also called false alarm), and (iii) False negative (FN) representing the number of pedestrians that are not detected, but they are in the GT (also called detection miss). Hence, each reported/detected bounded box is either true positive or false positive, and each ground truth bounding box is either true positive or false negative. So, there are no true negatives.

In general, MPDs apply some constraints on the reported bounding boxes to improve the pedestrian detection performance (i.e. accuracy and precision). The two most common constraints are the bounding box area/size and the bounding box confidence score. Normally, MPDs apply a fixed lower area threshold on the reported bounding box, so detectors ignore any bounding box with area less than the predefined lower area threshold. In other words, detectors will consider only those pedestrian with area/size greater than the lower area threshold as true positive [4]- [6]. On the other hand, some pedestrian detectors apply a fixed confidence score threshold on the reported bounding boxes to improve the detect performance. Thus, the bounding boxes with confidence score greater than the fixed threshold would be reported, and the bounding boxes with confidence score lower than the fixed threshold would be removed/ignored.

In [13], CNN MPD is used to detect pedestrians in a given video frame, wherein the detected bounding boxes with confidence score greater than 0.5 are accepted as true positive. In [14], a fixed threshold for upper confidence is used to create a confidential detection set, wherein detections with low confidence scores are removed from the original detection set at first step. In [15], fixed thresholds for upper and lower confidence scores are used and a sparse optical flow filter is applied to enhance the quality of detections, wherein the upper and lower confidence score thresholds are fixed for all frames in a given video.

In contrast, applying a lower confidence threshold on the reported bounding boxes to detect all existing pedestrians in the video frames at the cost of increasing the number of false positive detections. This is the case for KDNT [8] and FRCNN [7] where all detected bounding boxes are reported. It should

(a) (b) (c)

Fig. 1: Three consecutive frames: (a) Frame 708, (b) Frame 709, and (c) Frame 710 as taken from the MOT17-05 video sequence. The KDNT [8] detector detects the same pedestrian with three different confidence scores in successive frames.

(a) (b) (c)

Fig. 2: Same as Figure 1 except for the detection of a different pedestrian using the FRCNN [7] detector. Three consecutive frames: (a) Frame 666, (b) Frame 667, and (c) Frame 668 as taken from the MOT17-05 video sequence. As was the case for the KDNT detector, the FRCNN detector detects the same pedestrian with three different confidence scores in successive frames.

be noted that the same person can appear very differently during its presence in a given video depending on the changes in the background, local illumination, contrast, etc. Thus, the same person may be detected with different confidence scores in two consecutive frames. Therefore, applying upper or lower confidence score thresholds is not a desirable a approach, because the threshold value may vary during a given video or over different videos. Furthermore, KDNT [8], FRCNN [7] and some other MPDs generate pedestrian detections for each frame independently, ignoring inter-frame relationships that exist between consecutive frames. It should be noted that if a pedestrian is present in a frame at time $t-1$ with a high confidence score it will most likely be present in the next frame at time t. For the purpose of illustration, Figure 1 shows that KDNT [8] detects the same pedestrian with three different confidence scores in three consecutive frames. Figure 2 shows similar example for the FRCNN [7] detector.

In this paper, we propose a novel and efficient real-time method to improve the performance of multiple pedestrian detector (MPD) by introducing post-processing steps. The proposed method is causal so it only uses information from the current frame and past frames. The proposed post-processing steps use an adaptive approach to determine both area and confidence score constraints, and these steps can work on the output of any multiple pedestrian detector. For this purpose, we

select four of the state-of-the-art MPDs, namely KDNT [8] , FRCNN [7], SDP [16] and DPM [17], based on the pedestrian detection challenge in multiple object tracking benchmark (MOT Challenge 2017) [11]. The main contributions of this paper are:

1) Remove outlier detections by using an adaptive area threshold for each frame. We calculate both the lower area threshold, denoted by θ_L, and the upper area threshold, denoted by θ_H, for each frame. Hence, we achieve a dynamic setup for both the lower and upper area thresholds.
2) Adapt a dynamic approach to determine both the upper confidence score threshold value, denoted by α_H, and the lower confidence score threshold value, denoted by α_L, for each frame. Hence, the upper and lower confidence score thresholds are not fixed and they vary with time for a given video sequence.
3) Impose the inter-frame relationship by propagating the high confidence pedestrian detections from the previous frame to the current frame. Motivated by the fact that if a pedestrian present in a frame at time $t-1$ with high confidence score it is most likely to be present in the next frame at time t.

II. Proposed Method

The proposed method is described in terms of the proposed post-processing steps. These post-processing steps enable MPDs to be more accurate, precise and tolerant to false positive detections in generating pedestrian detections. An adaptive approach has been used to set both area and confidence score constraints.

A. Post-Processing Step 1: Remove outliers in detections

MPD estimates the position and size of bounding box for of the detected pedestrians. To reduce the number of false positive, we calculate the area of the detected bounding boxes and analyze the area distribution in each frame. For frame at time t, the bounding boxes with associated area less than the lower area threshold, denoted by θ_L^t, will be removed. Also, the bounding boxes with associated area greater than the upper area threshold, denoted by θ_H^t, will be removed. To set the values for θ_L^t and θ_H^t threshold parameters, we adapt a dynamic approach. For each frame, we calculate both mean, denoted by μ_A, and standard deviation, denoted by σ_A, for the area distribution. Then, we eliminate any detected bounding box associated with area above $\theta_H = (\mu_A + 2\sigma_A)$, and any bounding box associated with area below $\theta_L = (\mu_A - 2\sigma_A)$. Hence, we are able to remove outlier pedestrian detections for each frame. It should be noted that around 95.45% of the sample data lie between $\theta_L = (\mu_A - 2\sigma_A)$ and $\theta_H = (\mu_A + 2\sigma_A)$. Let d_i denotes a detection in a given frame, $D^t = \{d_0, d_1, \cdots, d_N\}$ be the set of original detections in the frame at time t where N is the total number of detections, $D_A^t = \{d_0, d_1, \cdots, d_M\}$ be the set of detections in the frame at time t after running Step 1 where $M \leq N$, and $\textbf{Area}(d_i)$ be the area of d_i. So the conditional proposition with quantifier represents Step 1 is

$$\forall d_i \in D^t : \theta_L^t \leq \textbf{Area}(d_i) \leq \theta_H^t \rightarrow d_i \in D_A^t$$

B. Post-Processing Step 2: Propagate the high confidence pedestrian detections from previous frames

Motivated by the fact that if a pedestrian is present in a frame at time $t-1$ with a high confidence score it is most likely to have a pedestrian in the next frame at time t. As discussed earlier and shown in Figure 1 and 2, applying a fixed upper or lower confidence threshold is not a desirable approach, because a good threshold value may vary over different frames in a video. We adapt a dynamic approach to set the upper confidence threshold value for a frame at time t, which is denoted by α_H^t. In each frame, we analyze the distribution of the confidence scores for the detected bounding boxes, and we use the third quartile value as the upper confidence threshold. The third quartile, denoted by Q_3 , is the median of the upper half of the data set. So, 25% of the detected bounding boxes would have confidence scores more than $\alpha_H^t = Q_3$. Similarly, we adapt a dynamic approach to set the lower confidence threshold for the frame at time t, which is denoted by α_L^t. We use the first quartile as the low confidence threshold value. The first quartile, denoted by Q_1, is the median of the lower half of the data set. So, 75% of the detected bounding boxes would have confidence scores more than $\alpha_L^t = Q_1$. We follow Algorithm 1 to propagate the high confidence pedestrian detections from the previous frame, and create the final detection set for the current frame. In addition to containing the high confidence detections from the current frame, the final detection set also contains the high confidence detections from the previous frame that have significant overlaps with the low confidence detections in the current frame. It should be noted that we calculated the values of α_H^t and α_L^t for each frame in real-time.

Notation: Let $D_C^t = \{d_0, d_1, \cdots, d_R\}$ be the set of detections in the frame at time t where $R \leq M \leq N$ after executing Algorithm 1 , $D_{HL}^t = \{d_{hl_0}, d_{hl_1}, \cdots, d_{hl_Z}\}$ be the set of all detections in the frame at time t that have confidence score greater than α_L where $Z \leq R$, $D_{LL}^t = \{d_{ll_0}, d_{ll_1}, \cdots, d_{ll_W}\}$ be the set of all detections in the frame at time t that have confidence score less than α_L where $W \leq R$ and $Z+W = R$, $D_{HH}^t = \{d_{hh_0}, d_{hh_1}, \cdots, d_{hh_V}\}$ be the set of all detections in the frame at time t that have confidence score greater than α_H where $V \leq Z \leq R$, $\textbf{conf}(d_i)$ be the confidence score of d_i, and $\textbf{IOU}(d_x, d_y)$ denotes the intersection-over-union of the bounding boxes of detections d_x and d_y.

$$\textbf{IOU}(d_x, d_y) = \frac{|d_x \cap d_y|}{|d_x \cup d_y|}$$

III. Experimental details and Results

We evaluated and tested the proposed online method to improve the performance of a given person/pedestrian detector using the pedestrian detection challenge in multiple object tracking benchmark (MOT Challenge 2017) [11]. For this

ALGORITHM 1. PROPAGATE THE HIGH CONFIDENCE PEDESTRIAN DETECTIONS

Input: Detection set $D_A^t = \{d_0, d_1, \cdots, d_M\}$.
Output: Detection set $D_C^t = \{d_0, d_1, \cdots, d_R\}$.

Initialization: For the frame at $t = 0$

N1. Determine the $\alpha_L^0 = Q_1$ and $\alpha_H^0 = Q_3$ for confidence score distribution of D_A^0.

N2. At $t = 0$, all detections that have confidence scores greater then α_L are considered to be part of final detection set. Hence, 75% of the detected bounding boxes are included in the final detection set. So, the conditional proposition with quantifier is

$$\forall d_i \in D_A^0 : \textbf{conf}(d_i) \geq \alpha_L^0 \rightarrow d_i \in D_C^0$$

Loop Steps: Repeat Step L1 to L5 for each frame

L1. Determine the $\alpha_L^t = Q_1$ and $\alpha_H^t = Q_3$ for confidence score distribution of D_A^t.

L2. Build sets D_{HL}^t and D_{HH}^t. Also, start building D_C^t by using the following conditional proposition with quantifier

$$\forall d_i \in D_A^t : \textbf{conf}(d_i) \geq \alpha_L^t \rightarrow d_i \in D_{HL}^t$$
$$\forall d_i \in D_A^t : \textbf{conf}(d_i) \geq \alpha_L^t \rightarrow d_i \in D_C^t$$
$$\forall d_i \in D_A^t : \textbf{conf}(d_i) \geq \alpha_H^t \rightarrow d_i \in D_{HH}^t$$

L3. Build sets D_{LL}^t by using the following conditional proposition with quantifier

$$\forall d_i \quad \in D_A^t : \textbf{conf}(d_i) < \alpha_L^t \rightarrow d_i \in D_{LL}^t$$

L4. Propagate high confidence previous detections by using the following compound conditional proposition with nested quantifiers

$$\forall d_{ll_i} \in D_{LL}^t \forall d_{hh_j} \in D_{HH}^{t-1} : \Big((\textbf{IOU}(d_{ll_i}, d_{hh_j}) \geq 0.5) \rightarrow \Big(\forall d_{hl_k} \in D_{HL}^t : \textbf{IOU}(d_{hh_j}, d_{hl_k}) < 0.5 \rightarrow d_{ll_i} \in D_C^t \Big) \Big)$$

L5. Report D_C^t as final detections for the frame at time t.

purpose, we selected four of the state-of-the-art pedestrian detectors, namely KDNT [8], FRCNN [7], SDP [16] and DPM [17] based on the 2017 pedestrian detection MOT Challenge.

A. Evaluation metrics

We use the same quantitative evaluation criteria used in the MOT Challenge [11], [12]. In the pedestrian detection challenge, multiple metrics are used to evaluate the performance of any given pedestrian detectors. These metrics include multiple object detection accuracy (MODA), multiple object detection precision (MODP), average number of false alarms per frame (FAF), total number of true positives (TP), total number of false positives (FP), total number of false negatives (FN), precision, recall, and average precision (AP) taken over a set of reference recall values $(0 : 0.1 : 1)$.

(a) (b) (c)

Fig. 3: Part of the Frame-206 taken from the MOT17-04 video sequence. (a) Four bounding boxes detected by KDNT detector, (b) The outlier (wrong) bounding box has been correctly isolated by the proposed post-processing algorithm (Step 1 in Section II-A), (c) Three true positive bounding boxes are correctly retained after applying the proposed method. The complete video is available at
`https://youtu.be/CytR_WFF5ys`

B. Results

We applied the proposed post-processing steps in a sequential manner. For each frame at time t, we first performed the post-processing Step 1, then post-processing Step 2. Figure 3 shows part/region of the Frame-206 taken from the MOT17-04 video sequence. It can be seen from Figure 3(a) that the KDNT [8] detector reported four bounding boxes in this region where actually only three pedestrians exist. Figure 3(b) shows that the proposed post-processing Step 1 is able to identify the wrong bounding box as an outlier based on the area constraint. Figure 3(c) shows that three true positive bounding boxes are reported and the outlier bounding box is removed after applying the proposed method. Therefore, the proposed method reduces the number of false positive detections in this frame which leads to reduction in the number of false alarms per frame (FAF).

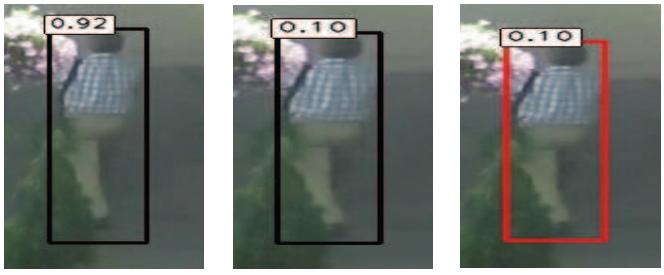
Figure 4 shows part/region from two consecutive frames (532 and 533) taken from the MOT17-04 video sequence. From these two frames, it can be seen that the FRCNN [7] detectors detect the same pedestrian with two different confidence scores. In Figure 4(a), the pedestrian was detected with a high confidence score of 0.92. However, in Figure 4(b), the same pedestrian was detected with a low confidence score of 0.10. Thus, the bounding box in Frame 533 is most likely to be removed after applying a confidence threshold constraint. This will lead to reduction in the number of true positive detections in this frame. Figure 4(c) shows that the proposed post-processing Step 2 is able to identify and recover the pedestrian in Frame 533 despite a low confidence score.

The proposed post-processing steps are tested on the video sequences taken from the pedestrian detection challenge– MOT Challenge 2017. We applied the proposed method on the four of the state-of-the-art pedestrian detectors, KDNT [8], FRCNN [7], SDP [16] and DPM [17]. The modified detectors are named as MKDNT (Modified-KDNT), MFRCNN (Modified-FRCNN), MSDP (Modified-SDP) and MDPM (Modified-DPM), respectively. Table I shows the

TABLE I: Performance results of the proposed methods, MKDNT, MFRCNN, MSDP and MDPM, and the four of the state-of-the-art pedestrian detectors, KDNT [8], FRCNN [7], SDP [16] and DPM [17], on the pedestrian detection challenge-MOT Challenge 207 (accessed on February 6, 2018).

MPD	AP↑	MODA↑	MODP↑	FAF↓	TP ↑	FP↓	FN↓	Precision↑	Recall ↑	Total Dets
KDNT [8]	0.89	67.10%	80.10%	4.80	105473	28623	9091	78.70%	92.10%	134096
MKDNT (our)	**0.89**	**75.90%**	**80.30%**	**2.70**	103143	**16185**	11421	**86.40%**	90.00%	119328
FRCNN [7]	0.72	68.50%	78.00%	1.70	88601	10081	25963	89.80%	77.30%	98682
MFRCNN (our)	0.71	**69.20%**	**78.40%**	**1.10**	86086	**6774**	28478	**92.70%**	75.10%	92860
SDP [16]	0.81	76.90%	78.00%	1.30	95699	7599	18865	92.60%	83.50%	103298
MSDP (our)	**0.81**	**77.40%**	**78.10%**	**0.80**	93738	**5024**	20826	**94.90%**	81.80%	98762
DPM [17]	0.61	31.20%	75.80%	7.10	78007	42308	36557	64.80%	68.10%	120315
MDPM (our)	**0.61**	**43.50%**	**76.00%**	**4.20**	74546	**24659**	40018	**75.10%**	65.10%	99205

Evaluation metrics with symbol (↑) indicates higher score is better; while for evaluation metrics with symbol (↓) indicates lower score is better



(a) (b) (c)

Fig. 4: Corresponding windows taken from two consecutive frames of the MOT17-04 video sequence: (a) Frame 532 and (b) Frame 533. (a) shows the output of the FRCNN [7] detector with a confidence score of 0.92, and (b) shows the output of the FRCNN detector for the same pedestrian with a confidence score of 0.10. (c) illustrates that the proposed post-processing algorithm (Step 2 in Section II-B), is able to identify and recover the pedestrian as a true positive detection even if it is detected with a low confidence score in the frame 533, as compared to frame 532. The complete video is available at `https://youtu.be/AynWszq7VZs`

quantitative evaluations for the performance of the proposed methods and the four of the pedestrian detectors, KDNT [8], FRCNN [7], SDP [16] and DPM [17]. Also, this comparison can be found in the MOT Challenge website, `https://motchallenge.net/results/MOT17Det/`.

The performance of the proposed MKDNT method is compared with the original KDNT [8] method. Table I shows that the proposed MKDNT achieves better MODA, MODP, FAF, FP, and precision as compared to KDNT. There were significant improvements in terms of MODA, FAF, FP, and precision metrics. MODA increases from 67.10% to 75.90% with an improvement of 8.80%. FAF decreases from 4.80 to 2.70 with an improvement of 2.10. FP decreases from 28623 to 16185 with an improvement of 12438. Precision increases from 78.70% to 86.40% with an improvement of 7.70%. It is noted that even though the proposed MKDNT method reduces the total number of detections from 134096 to 119328 (14768 detections are thrown out), it achieves better performance in terms of MODA, MODP, FAF, FP, and precision compare to KDNT. Moreover, the proposed MKDNT method provides a similar AP score to that of KDNT, despite fewer detections. The proposed MKDNT method achieves these improvements at the cost of decreasing TP by 2330 and increasing FN by 2330. It should be noted that these 2330 TP detections represent 16% from the total detections that are removed by the proposed MKDNT method.

The performance of the proposed MFRCNN method is also compared with the original FRCNN [7] method. Table I shows that the proposed MFRCNN achieves better MODA, MODP, FAF, FP, and precision as compared to FRCNN. MODA increases from 68.50% to 69.20% with an improvement of 0.70%. MODP increase from 78.0% to 78.40% with an improvement of 0.40%, FAF decreases from 1.70 to 1.10 with an improvement of 0.60. FP decreases from 10081 to 6774 with an improvement of 3307. Precision increases from 89.80% to 92.70% with an improvement of 2.90%. It is noted that even though the proposed MFRCNN method reduces the total number of detections from 114564 to 114564 (5822 detections are thrown out), it achieves better performance in terms of MODA, MODP, FAF, FP, and precision as compared to FRCNN. Moreover, the proposed MFRCNN method provides an almost similar AP score to that of FRCNN, even though it used a lesser number of detections. The proposed MFRCNN method achieves these improvements at the cost of decreasing TP by 2515 and increasing FN by 2515. It should be noted that these 2515 TP detections represent 43% of the total detections that are removed by the proposed MFRCNN method. Similarly, Table I shows that the proposed methods MSDP and MDPM achieve better MODA, MODP, FAF, FP, and precision as compared to SDP and DPM, respectively.

Table I shows that the proposed post-processing steps were able to significantly improve the performance of KDNT, FRCNN, SDP and DPM detection methods. It should be mentioned that the improvements were achieved without any extra training, fine-tuning or modification to the original detection methods. Furthermore, the proposed detection methods (MKDNT, MFRCNN, MSDP and MDPM) are more accurate, precise and tolerant to noise detections/false positive detections than the original detection approaches. Table I shows that the proposed method provides a higher improvement to the performance of KDNT and DPM than FRCNN and SDP. We believe that this is related to both how the detection method is trained and how the true positive detections are

reported. For instance, the confidence scores of the reported pedestrian detections from DPM [17] were between -0.5 and 4.8, whereas the confidence scores of the reported pedestrian detections from SDP [16] were between 0.4 and 1.0. It can be seen from Table 1 that KDNT and DPM reported more detections than FRCNN and SDP, which resulted in improved performance in terms of AP, TP, and FN.

Furthermore, the proposed post-processing steps work on the output of a given pedestrian detector, and do not generate any new pedestrian detections. Thus, if some pedestrians are not detected by the pedestrian detector then the proposed steps will not be able to recover these undetected pedestrians. It should be noted that the average precision (AP) score is considered strong evidence of improvement in terms of both recall and precision parameters. Table I shows that the proposed methods provide a similar or almost similar AP score to that of the selected detection methods.

IV. Conclusion

In this paper, we have proposed a novel and efficient online method to improve the performance of a given multiple pedestrian detector (MPD). New post-processing steps have been proposed to make the pedestrian detector more accurate, precise and tolerant to false positive detections in generating pedestrian detections. In the proposed post-processing steps, an adaptive approach has been used to set both of the area and confidence score constraints. For setting an adaptive area constraint, the area distribution of the detected bounding boxes in a given frame is analyzed, and then the upper and lower area threshold are dynamically determined to remove outlier detections. For setting an adaptive confidence constraint, the confidence scores of the detected bounding boxes in a given frame are sorted, and then the upper and lower confidence thresholds are dynamically determined by using first and third quartiles, respectively. In order to study the performance of the proposed method, four of the-state-of-the-art pedestrian detectors, KDNT [8], FRCNN [7], SDP [16] and DPM [17], were selected. The results have shown that the proposed method significantly improves the performance of the selected detection methods without any extra training, fine-tuning or modification to the original detection methods. Furthermore, it has been shown that considering the inter-frame relationship is essential to improve the performance of a given pedestrian detector in video surveillance applications. Finally, it should be noted that the proposed post-processing steps work on the output of the pedestrian detectors, and these steps are not limited to a specific pedestrian detector algorithm

Acknowledgment

This research is funded in part by Natural Science and Engineering Research Council (NSERC), Canada through the collaborative research and development grant CRDPJ $461882-13$. Computations were performed on the SOSCIP Consortiums [Blue Gene/Q, Cloud Data Analytics, Agile and/or Large Memory System] computing platform(s). SOSCIP is funded by the Federal Economic Development Agency of Southern Ontario, the Province of Ontario, IBM Canada Ltd., Ontario Centres of Excellence, Mitacs and 15 Ontario academic member institutions in Canada

References

[1] A. W. M. 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, July 2014.

[2] Xi Li, Weiming Hu, Chunhua Shen, Zhongfei Zhang, Anthony Dick, and Anton Van Den Hengel, "A Survey of Appearance Models in Visual Object Tracking," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 4, pp. 58:1–58:48, Oct. 2013.

[3] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: A benchmark," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 304–311.

[4] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.

[5] Jan Hosang, Mohamed Omran, Rodrigo Benenson, and Bernt Schiele, "Taking a deeper look at pedestrians," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4073–4082.

[6] Y. Tian, P. Luo, X. Wang, and X. Tang, "Deep Learning Strong Parts for Pedestrian Detection," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 1904–1912.

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in Neural Information Processing Systems*, pp. 91–99. Curran Associates, Inc., 2015.

[8] Fengwei Yu, Wenbo Li, Quanquan Li, Yu Liu, Xiaohua Shi, and Junjie Yan, "POI: Multiple Object Tracking with High Performance Detection and Appearance Feature," in *European Conference on Computer Vision*. Springer, 2016, pp. 36–42.

[9] Z. Cai, M. Saberian, and N. Vasconcelos, "Learning Complexity-Aware Cascades for Deep Pedestrian Detection," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec. 2015, pp. 3361–3369.

[10] E. Bochinski, V. Eiselein, and T. Sikora, "Training a convolutional neural network for multi-class object detection using solely virtual world data," in *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Aug. 2016, pp. 278–285.

[11] Anton Milan, Laura Leal-Taixe, Ian Reid, Stefan Roth, and Konrad Schindler, "MOT16: A Benchmark for Multi-Object Tracking," *arXiv:1603.00831 [cs]*, Mar. 2016.

[12] Rainer Stiefelhagen, Keni Bernardin, Rachel Bowers, John Garofolo, Djamel Mostefa, and Padmanabhan Soundararajan, "The CLEAR 2006 Evaluation," in *Multimodal Technologies for Perception of Humans*. Apr. 2006, Lecture Notes in Computer Science, pp. 1–44, Springer, Berlin, Heidelberg.

[13] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*, Sept. 2016, pp. 3464–3468.

[14] J. Chen, H. Sheng, Y. Zhang, and Z. Xiong, "Enhancing Detection Model for Multiple Hypothesis Tracking," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 2143–2152.

[15] V. Eiselein, E. Bochinski, and T. Sikora, "Assessing post-detection filters for a generic pedestrian detector in a tracking-by-detection scheme," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Aug. 2017, pp. 1–6.

[16] F. Yang, W. Choi, and Y. Lin, "Exploit All the Layers: Fast and Accurate CNN Object Detector with Scale Dependent Pooling and Cascaded Rejection Classifiers," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 2129–2137.

[17] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, Sept. 2010.