Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 15, no. 6, June 2019, pp. 1538-1547
DOI: 10.23940/ijpe.19.06.p5.15381547

Method based on Separation Confidence Computation and Scale Synthesis Optimization for Real-Time Target Detection in Streetscape Videos

Jianmin Liu[a,b,*], Minhua Yang[b], and Jianmei Tan[a]

[a]*School of Information and Statistics, Guangxi University of Finance and Economics, Nanning, 530003, China*
[b]*School of Geosciences and Info-Physics, Central South University, Changsha, 410000, China*

Abstract

This study proposes a method for the real-time detection and recognition of targets in streetscape videos. The proposed method is based on separation confidence computation and scale synthesis optimization. First, on the basis of generalization in transfer learning, we combine a fine-tuning method suitable for non-convex optimization and adaptive moment estimation in high-dimensional space. Then, we dynamically adjust the learning rates of parameters on the basis of first and second gradient moment estimations. We establish the framework and implementation steps of the proposed method by organically combining regular term super-parameter generalization and hard-example mining technology. We use the proposed method to detect and recognize targets in streetscape videos with high frame rates and high definition. Furthermore, we experimentally demonstrate that the accuracy and robustness of our proposed method are superior to those of conventional methods.

Keywords: object detection; separation confidence computation; scale synthesis optimization; transfer learning; streetscape videos

(Submitted on March 20, 2019; Revised on April 8, 2019; Accepted on June 6, 2019)

1. Introduction

Numerous Internet companies provide online streetscape image, mapping, and navigation services. All types of network media store large numbers of streetscape images and have gradually accumulated a massive volume of streetscape images and video data over time. The basic hardware conditions and data sources necessary for the automatic detection of related targets in videos are currently available. Moreover, accurate target detection is the basis of subsequent classification, target tracking, and behavioral analysis [1-2].

For automated streetscape video target real-time detection recognition techniques, it is key to improve the intelligence level of monitoring system networks. By applying the depth learning method to this field, multiple video sources and data sources are aggregated to realize intelligent detection and recognition for the related departments. It is very meaningful to provide accurate and real-time information for decision making.

The deep learning algorithm was first applied in image classification. Krizhevsky et al. proposed AlexNet, a deep convolution neural network that comprises eight learning layers. Specifically, AlexNet comprises five convolution layers and three fully connected layers. It achieved a top-five classification error rate as low as 15.3% when applied to classify images from the ImageNet dataset on the basis of dropout and random gradient descent [3]. In 2015, He proposed the use of deep residual neural networks to address the performance degradation caused by increasing the number of layers of the deep neural network to 152. To facilitate training, they used congruent mapping to directly connect the preceding output layer to the upper layer. The classification error rate of the deep residual neural network when used to classify images form the ImageNet dataset decreased to 3.6%, which is 5% less than the error rate achieved through human visual detection [4].

* Corresponding author.
E-mail address: jianminliu2007@163.com

In 2013, Szegedy et al. proposed a target detection method based on convolution neural networks and regression. This detection method yielded a mAP of 30.5% when applied on the VOC2007 dataset [5]. In 2014, Girshick et al. presented a region-based convolution neural network (RCNN), which transforms the target detection problem into a classification problem solvable by the convolution neural network. The mAP of the RCNN reached 58% when used to classify images from the VOC2007 dataset [6]. In 2015, Girshick et al. developed a region-based fast convolution neural network (FAST RCNN) that maps regions directly to the feature graph on the last convolution layer of the convolution neural network. FAST RCNN increased the computational speed by transforming the target detection problem into a classification problem solvable by the convolution neural network. It yielded a mAP of 68% when applied on the VOC2007 dataset [7].

In 2015, Girshick and He proposed a region-based faster convolution neural network (Faster-RCNN), which aggregated feature extraction, proposal extraction, bounding box regression, and classification to detect targets with quasi real-time speed. The mAP of VGG-based Faster-RCNN reached 76.4% when used to classify images from the VOC2007 dataset [8]. In 2016, Redmon developed a target detection algorithm that was based on the regression method. The algorithm achieved a mAP of 57.9% in the VOC 2012 dataset on the basis of end-to-end network from image input to output target location and the detected category [9].

Image classification and target detection with the deep learning neural network framework has become an important development model [10]. However, some methods for image detection and classification have been unable to keep up with the rapid development of intelligent monitoring, intelligent traffic, and other fields, as well as the expansion of data sources for streetscape videos with high frame rate, definition, and complex real scenes [11-12]. Therefore, a new method needs to be studied for the real-time detection and recognition of targets in streetscape videos.

2. Method

As shown below, the proposed method is based on separation confidence computation and scale synthesis optimization. First, on the basis of generalization in transfer learning, we combine a fine-tuning method suitable for non-convex optimization and adaptive moment estimation in a high-dimensional space. Then, we dynamically adjust the learning rates of parameters on the basis of first and second gradient moment estimations. We establish the framework and implementation steps of the proposed method by organically combining regular term super parameter generalization and hard-example mining technology. Through the organic combination of the above methods, we use the proposed method to detect and recognize targets in streetscape videos with high frame rates and high definition.

2.1. Separation Confidence Computation

In contrast to the selective target detection algorithm [13], image scanning generates rectangular selective boxes [14-15]. This behavior improves real-time performance. In image scanning, the video frame is divided into a fixed number of rectangular selection frames that may contain the object of interest. The corresponding target and classification probability are calculated based on the area encompassed by the rectangular selection frame. Thus, the calculation cost of this method is considerably reduced. The specific steps are as follows: first, the length and width of the video input image are divided into N parts, and the image is divided into $N \times N$ segments. The target object is predicted from the segment containing the center of the object. Videos of streetscapes, such as public squares, pedestrian streets, and dense traffic arteries, contain high densities of images of pedestrians and vehicles. To avoid missing valuable targets, N is set as 9, 11, and 13. N is set as 9, 11, and 13 in accordance with theoretical analysis, as shown in Figures 1-3. The algorithms are trained and tested independently, and the mAP [16] and detection rate are calculated.

Figure 1. Horizontal–vertical division of a streetscape video image frame (N = 9)

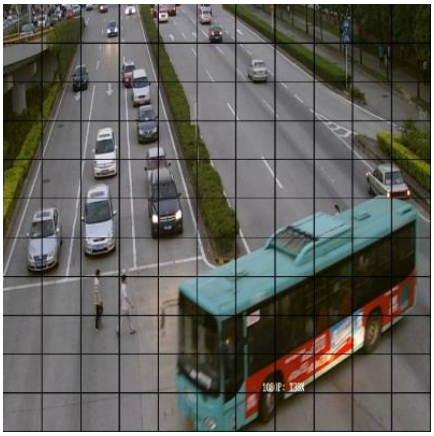

Figure 2. Horizontal-vertical division of a streetscape video image frame (N = 11)

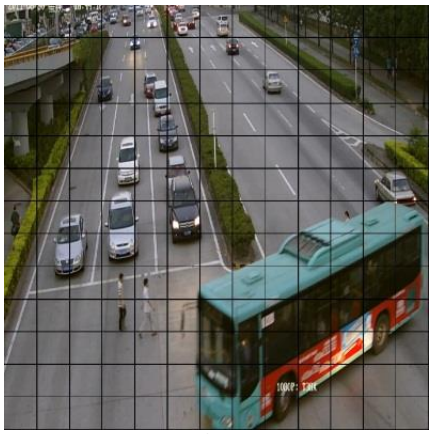

Figure 3. Horizontal-vertical division of a streetscape video image frame (N = 13)

To enhance the generalization ability and to account for the different length-width ratios of different types of targets and the variations in the length-width ratios of the same type of targets, we designed six kinds of anchor selection boxes with different proportions. We specifically designed these boxes to account for pedestrians, non-motor vehicles, motorcycles, cars, truck vehicles, and buses in images and to increase the probability that the anchor selection boxes contain the targets of interest. The length-width ratios of the six types of anchor selection boxes are designed as 1:1, 1:2, 2:1, 2:2, 2:4, and 4:2 (Figure 4).

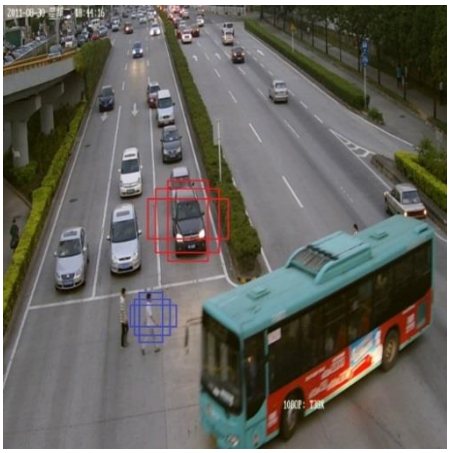

Figure 4. Examples of six selection boxes in image frames taken from a streetscape video

As shown in Figure 4, after each image frame of the streetscape video is processed, we set N = 9 and B = 6. Pedestrians, various non-motor vehicles, and motor vehicles in the video are set as the detection targets. We set six categories, including pedestrians, non-motor vehicles, motorcycles, cars, truck vehicles, and buses. The output is the detector.

Taking $N = 9$, the image is divided into $9 \times 9 = 81$ segments. Each segment is used to predict six categories with six anchor selection boxes. The six confidence levels are represented by 0, 1, 2, 3, 4, and 5. The coordinates of each anchor selection box are [a b w h], and each selection box has a confidence level. Each segment is used to predict $6 + 6 \times 4 + 6 = 36$ parameters. Each image frame from the video predicts $9 \times 9 \times 36 = 2916$ parameters. Taking $N = 11$, the image is divided into $11 \times 11 = 121$ segments. Each segment contains five categories and provides six selection boxes. The six confidence levels are represented by 0, 1, 2, 3, 4, and 5. Furthermore, the coordinates of each selection box are [a b w h]. Each selection box has a confidence level. Thus, each segment is used to predict $6 + 6 \times 4 + 6 = 36$ parameters. A total of $11 \times 11 \times 36 = 4356$ parameters are predicted in one image. We set $N = 13$. Thus, the image is divided into $13 \times 13 = 169$ segments. Each segment contains five categories and six selection boxes. The six confidence levels are represented by 0, 1, 2, 3, 4, and 5. The coordinates of each selection box are [a b w h]. Each selection box has a confidence level. That is, each segment is used to predict $6 + 6 \times 4 + 6 = 36$ parameters. A total of $13 \times 13 \times 36 = 6084$ parameters are predicted per image. In the above three cases, the coordinates of each selection box are [a b w h]. The coordinate [a b] is normalized to [0–1] by the corresponding separation relative to the deviation of each frame image, and [w h] is normalized to [0–1] on the basis of the width and height of the image.

2.2. Competition of Scale-based Synthesis Optimization

In the detection of streetscape targets, such as traffic tools and pedestrians, two vectors are obtained: the predicted rectangular selection box that may contain the target and the value obtained after the classifier is applied to the rectangular selection box that may contain the target. However, mutual inclusions or overlapping may exist, thus occluding objects in the rectangular selection box. Thus, the appropriate strategies for synthesis optimization must be designed to resolve the problem of occlusion. Non-maximum suppression (NMS) screens elements and only allows the retention of the largest element [17] in an interval. This interval, in turn, contains two variables: the dimension and the range of the interval. Therefore, in this study, the scale-based synthesis optimization strategy is applied to identify the optimal rectangular selection box.

Given the influence of occlusion and position, low-scoring rectangular selection boxes may match with the relevant target. Therefore, although the initial rectangular selection box that surrounds each target is redundant, it can also be used to improve the accuracy of NMS before it is deleted. In this section, we discuss the method for the generation of the optimal rectangular selection box through NMS with scale-based optimization. When the rectangular selection box window W_u with the highest score (its detection score is S_u) is fully included in the rectangular selection box W_i with a low target score and large scale (its detection score S_i), and the detection scores $\frac{S_u - S_i}{S_u} < \lambda$ are met, W_i is retained and $S_u = S_i$ is implemented. Otherwise, W_i is directly deleted, and NMS is generated on the basis of the scale optimization strategy, which has strong robustness. Setting $\lambda \in [0.15, 0.2]$ considerably decreases the missed detection rate and false detection rate of the target rectangular selection box.

2.3. Model Generalization based on Transfer Learning

Although streetscape videos are easily obtained, the number of mature datasets and the amount of data that can be used for streetscape object label training remain limited. The inappropriate selection of the initialization mode of the network model parameters is an important reason for model overfitting. The relative scarcity of accurately tagged streetscape image data is a major technical problem encountered in the development of detection models. To solve this problem, a model generalization technology based on transfer learning [18] is introduced to transfer the pretraining network model to the application of streetscape classification.

On the basis of transfer learning technology, under class A domain D_A, task T_A, and $T_A \in D_A$, the pretraining knowledge optimization model M_B is transferred to the new application scene. We continue to use the network parameters and topology of the pretraining knowledge optimization model M_B, that is, we pretrain the algorithm with a massive, universal image dataset. Then, distinct areas in small-scale streetscape datasets are fine-tuned to improve the classification accuracy of the proposed method for specific applications.

2.4. Fine-Tuning of Stochastic Optimization based on Adaptive Moment Estimation

Stochastic optimization based on adaptive moment estimation is a weight-adjustment method for the optimization of non-

convex and high-dimensional space. The first moment estimation of the gradient is set to be m_t, and the second moment estimation of the gradient is set to be n_t. After continuously optimizing the learning rate by m_t and n_t, they can be regarded as the estimation of expectation $E|g_t|\ E|g_t^2|$. $\hat{m}_t\ \hat{n}_t$ is the correction of $m_t\ n_t$, which can be approximated as the unbiased estimation of the expectation, as shown in Equation (1):

$$\Delta\theta_t = -\frac{\widehat{m}_t}{\epsilon+\sqrt{\widehat{n}_t}} \times \eta \tag{1}$$

To achieve intensive resource demand, stochastic optimization based on adaptive moment estimation exploits the ability of the adaptive adjustment learning rate gradient method for high-performance computer sparse gradient and the adaptive learning rate gradient method for high-performance response to unsteady objects. The corresponding parameters are autonomously optimized to match their learning rates. In this way, the floating range of the stochastic optimization learning rate of the adaptive moment estimation is limited to $-\frac{\hat{m}_t}{\varepsilon+\sqrt{\hat{n}_t}}$. The parameters under this constraint can be applied to non-convex optimization because of their stability, which is suitable for large datasets and high-dimensional space.

2.5. Hard-Example Mining

A large number of samples should be input during network training to ensure effective training. In addition, each sample has different effects on the fine-tuning result of the network. Some good samples aid training, whereas hard examples complicate training. However, if the training set contains an excessive proportion of good examples, the generalization ability and accuracy of the network will be affected. The addition of h numbers of hard examples in the training set can improve the effectiveness of the whole network but also increases the computational time. After the completion of network training, new samples need to be added to the training set. We use online hard-example mining [19] to improve efficiency and accuracy rate.

2.6. Design of Compound Loss Function

We use a 9×9 segment as an example to design a compound loss function. In reference to the regularization term λ, the probability and the error between the predictive rectangular selection box and the real value are weighted to generalize the detection and recognition ability.

Input: $S_i(x_g, y_g, \omega_g, h_g)$ refers to the labeled target selection box, and x_g, y_g refers to the central point.

Input: $S\ (x, y, \omega, h)$ refers to the preselection box of the target detection algorithm, and x, y refers to the central point.

I: The number of segments in each image frame. Each segment contains six selection boxes (i = 0....max = $N \times N$, $9 \times 9 \times 6$).

$\lambda_{noobject}$: The weight of non-target separation IoU loss for the gradient computation is set to 0.1 for optimization.

λ is set to 0.2 to optimize the weight of the target separation IoU loss and classification loss for the gradient computation.

$\prod_i^{object}$ represents a selection box of the separation divided by each frame image. The tab is i (i = 0....485, $9 \times 9 \times 6$) and contains the concerned target.

$\prod_i^{noobject}$ represents a segment in the selection box divided by each frame image. The tab is i (i = 0....486, $9 \times 9 \times 6$) and excludes the target output of concern.

Given that 1 out of 81 segments is required to calculate the probabilities of six categories, the original data of each input generate 486 probabilities through the network. The probability of a space point is set to $Pr(Object)$. Then, the target can

be quantitatively measured. After 1 out of 81 segments is given by six categories of probabilities, the probability that the segment contains any target $Pr(Object)$ must be quantified. Then, after each segment is divided into six categories, a single independent category of unconditional probability compounds the categorization probability of a target after identification, as shown in Equation (2):

$$Pr(Car) = Pr(Object) \times Pr(Car|Object) \tag{2}$$

To summarize, the weighted linear combination of the independent loss function is used to optimize $Loss_{all}$, the total loss value generated, as shown in Equation (3):

$$\begin{aligned} Loss_{all} &= l_2\ loss + IoU\ \ loss + Class\ loss \\ &= \sum_{i=0}^{80}\left[\prod_i^{object}((x_i - x_g)^2 + (y_i - y_g))^2\right] \\ &+ \sum_{i=0}^{80}\left[\prod_i^{object}((\sqrt{\omega_i} - \sqrt{\omega_g})^2 + (\sqrt{h_i} - \sqrt{h_g}))^2\right] \\ &+ \lambda_{noobject}\sum_{i=0}^{80}\prod_i^{noobject}[ln(\frac{S_i \cap S_g}{S_i \cup S_g - S_i \cap S_g})]^2 \\ &+ \lambda\sum_{i=0}^{80}\prod_i^{object}[ln(\frac{S \cap S_g}{S_i \cup S_g - S_i \cap S_g})]^2 \\ &+ \lambda\sum_{i=0}^{80}\prod_i^{object}\sum_{c\in classifications}(p_i(c) - p_i(c_g))^2 \end{aligned} \tag{3}$$

2.7. Model Design

The designed model for the detection of targets in streetscape videos should achieve end-to-end image feature extraction and target detection and recognition. The training and testing of the model are shown in Figures 5 and 6. We design the whole convolution network architecture to enhance the adaptability of input data and fully exploit the spatial information of the target.

1) A fine-tuning method for model generalization based on transfer learning and stochastic optimization for adaptive moment estimation is adopted.

With the initial min-batch gradient descent of the model, every time of the iterative learning rate lr is changed from 0.01 to 0.001. This learning rate attenuation strategy aims to learn new knowledge without completely forgetting old knowledge.

2) The anchor selection box is competitively optimized through scale-based synthesis optimization.

When the model is trained for a specific input image, if the boundary of the anchor selection box exceeds the image boundary, training loss should be unaffected by such an anchor selection box. The loss value of this type of anchor selection box will be directly shielded. After shielding the anchor selection box that intersects with the image boundary, the remaining anchor selection boxes will overlap with each other at numerous regions. Therefore, the scale-based synthesis optimization strategy is used to select the anchor selection box for the development of competitive optimization with IoU > 0.7, and the anchor selection box that wins the competition is retained. Similarly, in the final output layer, at the confidence value > *P* and IoU > 0.7, each rectangular selection box is competitively optimized through the scale-based comprehensive optimization strategy, and the final winner is output.

3) The design of the whole convolution network architecture can enhance adaptability to the input data and fully exploit the spatial information of the target of concern. Enhancing data processing by using the target detection training dataset can appropriately expand the dataset size and reduce the overfitting.

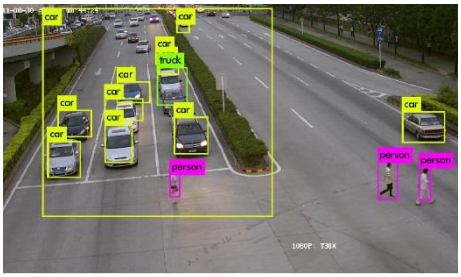

Figure 5. The proposed method for Streetscape video with D5 resolution 1 and instantaneous test output with $N = 9$

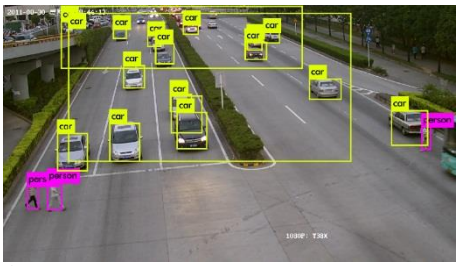

Figure 6. The proposed method for Streetscape video with D5 resolution 1 and instantaneous test output with $N = 11$

3. Experiments

3.1. Experimental Setup

The Pascal VOC 2007+2012 dataset [20], KITTI dataset [21], and Udacity dataset [22] discussed in Section 2.5 are adopted for pretraining and evaluating the proposed method for real-time target detection and recognition in streetscape videos. Sixteen streetscape videos with different resolutions are obtained from a traffic intersection monitoring terminal. Among these videos, nine are in D5 format. TargetLabelingInVideos tool is used to tag 319 video frames. The dataset is converted to the targeted fine-tuning training set with the same standard for transfer learning by a data preprocessing program, and 20% of data are reserved for testing data. We quantify the performance of the proposed method on the basis of the target detection comprehensive index mAP value [16]. This comprehensive evaluation index is obtained and calculated on the basis of recall, precision, and average precision (AP) and enables comprehensive measurement. To facilitate comparison with other methods, we define IOU > 0.5 as successful matching. The experimental platform is four-way 1080Ti and i7 7700K with a memory capacity of 4TB.

3.2. Analysis and Discussion of Experimental Results

We set $N = 9, 11, 13$ segments and perform independent training and testing. The instantaneous test output for streetscape videos with $N = 9, 11, 13$ and D5 resolution is shown as Figures 5-7. The obtained mAP and FPS are shown in Figures 8-10. As shown in Tables 1 and 2, the experiment involves 30 training cycles and a total of 95.6h of training time.

Table 1 shows the comparison of the mAP values of this method with those of other detection methods. Table 2 shows the comparison of the mAP and FPS of this method with those of other detection methods. mAP values tend to stabilize after the 25th training cycle when $N = 9, 11$, and 13. During 25 training cycles, the loss_bbox(train), loss_coverage(train), loss_bbox(val), and loss_coverage(val) values gradually converge and stabilize. The N values do not significantly affect the training and convergence results. The mAP is 90.9 and fps is 70 when $N = 9$. The mAP is 91.74 and fps is 57 when $N = 11$. The mAP is 91.56 and fps is 48. The above fps values are obtained from the detection of videos with D5 resolution.

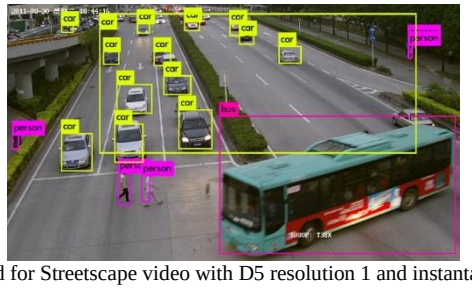

Figure 7. The proposed method for Streetscape video with D5 resolution 1 and instantaneous test output with $N = 13$

When the N value is changed from 9 to 11–13, the mAP value of pedestrian detection increases from 91.1 to 93.3–93.5, and the mAP value of non-motor vehicle detection changes from 90.5 to 92.4–92.55. Given the small size of pedestrians, non-motor vehicles, and other targets, segmentation intensity is positively related with accuracy.

The classification accuracy of the proposed method for motorcycles, cars, and other medium-sized targets is 89.8 and 91.7 when $N = 9$. The classification accuracy of the proposed method for motorcycles, cars, and other medium-sized targets is 90.2 and 91.6 when $N = 11$. The classification accuracy of the proposed method for motorcycles, cars, and other medium-sized targets is 90 and 91.1 when $N = 13$.

When $N = 11$, segmentation intensity is moderate, and the highest classification accuracy for motorcycles, cars, and other medium-sized targets is obtained. The classification accuracy of the proposed method for buses, trucks, and other large targets is 91.5 and 91.3 when $N = 9$. The classification accuracy of the proposed method for buses, trucks, and other large targets is 91.3 and 91.1 when $N = 11$. The classification accuracy of the proposed method for buses, trucks, and other large targets is 90.7 and 90.7 when $N = 13$. However, when $N = 9$, the intensity of separation weakens, and the classification accuracy for buses, trucks, and other large-sized targets increases.

As N is increased from 9 to 11–13, the computational cost increases, whereas the processing capacity decreases. Fps gradually decreases from 70 from 57–48. The processing capacity when $N = 9$ fully meets the requirement for target detection in streetscape videos shot by cameras with a high frame rate of 60 fps and can be also applied to multichannel videos and some platforms with insufficient processing capacity.

The processing capacity when $N = 11$ can fully meet the requirements for processing streetscape videos taken by a conventional 30 fps streetscape camera, but is insufficient for processing streetscape videos shot by a 60 fps cameras. When $N = 13$, the computational cost increases, and the processing capacity cannot meet the requirements for target detection from streetscape videos with a high frame rates of 60 fps. Nevertheless, the detection accuracy for small and distant targets in streetscape videos increases. Thus, N can be set to 13 for the accurate real-time detection of small targets from streetscape videos with low frame rates.

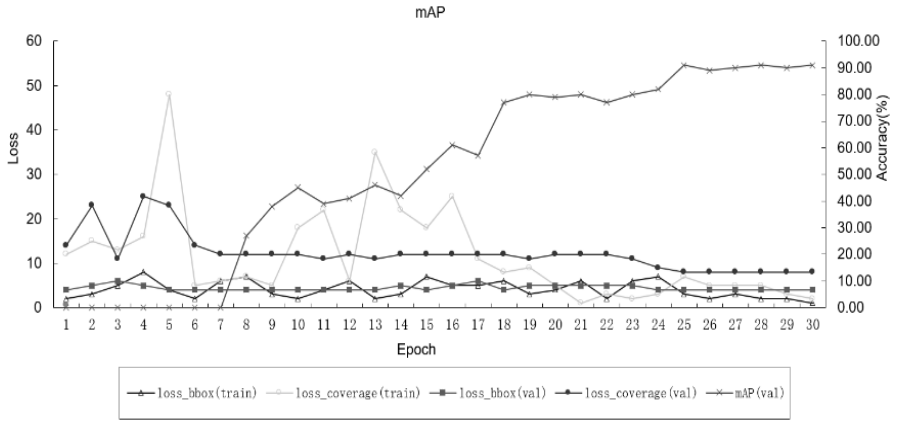

Figure 8. Training loss value and accuracy curves of the proposed method with 30 training cycles and $N = 9$

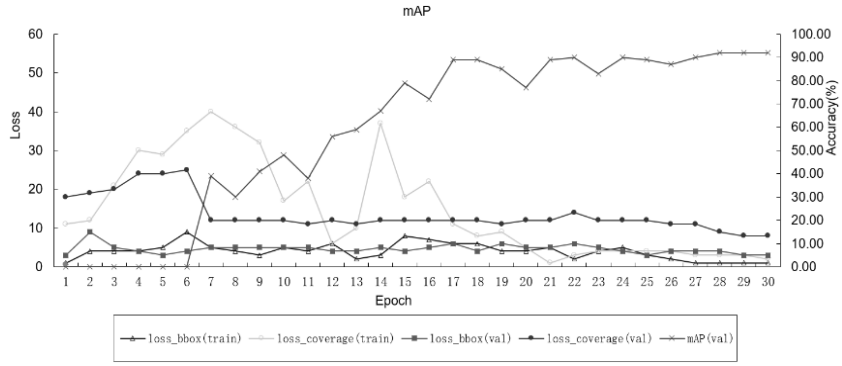

Figure 9. Training loss value and accuracy curves of the proposed method with 30 training cycles and $N = 11$

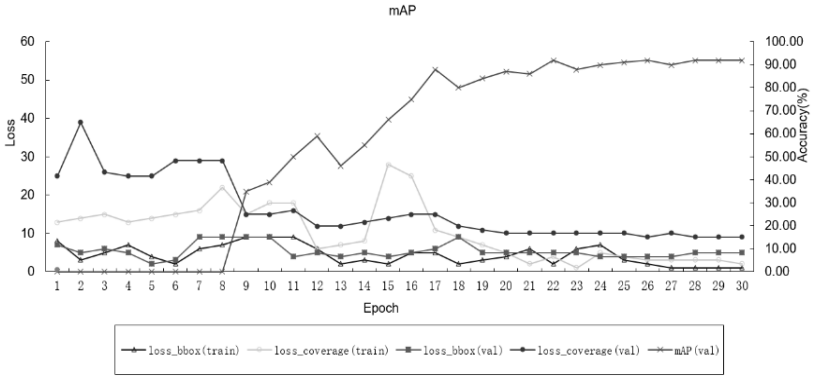

Figure 10. Training loss value and accuracy curves of the proposed method with 30 training cycles and N = 13

Table 1. Comparison of the mAP values of this method

Method	mAP	Person	Motor	Car	Bicycle	Bus	Truck
Faster R-RCN [8]	70.4	79.6	80.9	75.9	-	-	-
YOLO v2 [14]	73.4	81.3	83.4	76.5	-	79.8	-
ResNet [4]	80.7	86.5	84.8	76.6	-	78.6	-
DMP SVM [23]	36.1	43.1	39.	33.1	40.2	25.1	29.5
N = 9, D5 resolution	90.90	91.1	89.8	91.7	90.5	91.5	91.3
N = 11, D5 resolution	91.74	93.3	90.2	91.6	92.4	91.3	91.1
N = 13, D5 resolution	91.56	93.5	90	91.1	92.5	90.7	90.7

Table 2. Comparison of the mAP and FPS values of the proposed method with those of other detection methods

Method	mAP	Training time (hours)	FPS (testing)
Faster R-RCN [8]	70.4	179.5	7
YOLO v2 [14]	73.4	99.5	40
DMP SVM [23]	36.1	52.1	85
N = 9, D5 resolution	90.9	65.3	70
N = 11, D5 resolution	91.74	80.9	57
N = 13, D5 resolution	91.56	95.6	48

As shown in Tables 1 and 2, the method developed in this work integrates various methods for target detection. The proposed method yields a mAP of 90.9 and FPS of 70 when N = 9 for an input data source with D5 resolution. When N = 11, the proposed method obtains mAP and FPS values of 91.74 and 57, respectively. When N = 13, the proposed method obtains mAP and FPS values of 91.56 and 48, respectively. The detection accuracy and speed of the proposed method are superior to those of several classical methods, such as the Faster R-CNN ResNet, Faster R-CNN VGG-16, and YOLO v2 544×544 methods.

4. Conclusions

We proposed a real-time detection and recognition method based on separation confidence computation and scale synthesis optimization to address the problems encountered in target detection from streetscape videos with high frame rate and high definition. Moreover, we established the framework and implementation steps of the proposed method. mAP and FPS measurements indicate that the performance of the proposed method has been improved by the combination of regular term superparameter generalization and hard-example mining technology. The proposed method meets the technical requirements for the detection of targets in streetscape videos with high frame rate and high definition.

Acknowledgments

This work was supported by the Ph.D. Research Foundation of Guangxi University of Finance and Economics, Key Research Projects of Hunan Provincial Department of Education (No. 17A108), and Guangxi Natural Science Foundation.

References

1. S. Wu, D. Chen, and X. Wang, "Moving Target Detection based on Improved Three Frame Difference and Visual Background Extractor," in *Proceedings of 2017 10th International Congress on Image and Signal Processing, Biomedical Engineering and*

Informatics, IEEE, 2017
2. X. Shen, Z. Song, and H. Fan, "Data Level Moving Target Detection Algorithm based on Bernoulli Random Finite Set," *Iet Signal Processing*, Vol. 12, No. 6, 2018
3. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems*, pp. 1097-1105, 2012
4. K. He, X. Zhang, and S. Ren, "Deep Residual Learning for Image Recognition," in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016
5. S. Christian, T. Alexander, and E. Dumitru, "Deep Neural Networks for Object Detection," in *Proceedings of Conference on Neural Information Processing Systems*, pp. 2553-2561, 2013
6. R. Girshick, J. Donahue, and T. Darrell, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," arXiv Preprint arXiv: 1311.2524v3, 2014
7. R. Girshick, "Fast R-CNN," arXiv Preprint arXiv: 1504.08083, 2015
8. K. R. He and R. Girshick, "Faster R-Cnn: Towards Real-Time Object Detection with Region Proposal Networks," in Proceedings of the 28th International Conference on Neural Information Processing Systems, Vol. 1, pp. 91-99, 2016
9. J. Redmon, S. Divvala, and R. Girshick, "You Only Look Once: Unified, Real-Time Object Detection," arXiv Preprint arXiv: 1506. 02640, 2016
10. C. Szegedy, W. Liu, and Y. Jia, "Going Deeper with Convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-9, 2015
11. J. Li, H. C. Wong, and S. L. Lo, "Multiple Object Detection by Deformable Part-based Model and R-CNN," *IEEE Signal Processing Letters*, Vol. 1, No. 1, pp. 99, 2018
12. P. Dong and W. Wang, "Better, Region Proposals for Pedestrian Detection with R-CNN," in *Proceedings of Conference on Visual Communications and Image Processing*, pp. 1-4, IEEE, 2017
13. J. R. R. Uijlings, K. Sande, and T. Gevers, "Selective Search for Object Recognition," *International Journal of Computer Vision*, Vol. 104, No. 2, pp. 154-171, 2013
14. J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," arXiv Preprint arXiv: 1612.08242, 2016
15. W. Liu, D. Anguelov, and D. Erhan, "SSD: Single Shot MultiBox Detector," in *Proceedings of European Conference on Computer Vision*, Springer, Cham, pp. 21-37, 2016
16. M. Zhu, "Recall, Precision and Average Precision," Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, pp. 2-30, 2004
17. A. Neubeck and G. L. Van, "Efficient Non-Maximum Suppression," in *Proceedings of ICPR 2006 18th International Conference on Pattern Recognition*, pp. 850-855, IEEE, 2006
18. S. J. Pan and Q. Yang, "A Survey on Transfer Learning," *IEEE Transactions on Knowledge & Data Engineering*, Vol. 22, No. 10, pp. 1345-1359, 2010
19. A. Shrivastava, A. Gupta, and R. Girshick, "Training Region-based Object Detectors with Online Hard Example Mining," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 761-769, 2016
20. "The Pascal Visual Object Classes Challenge 2012 (voc2012) Results (2012)," (http://www.pascal-network.org /challenges/VOC/, last accessed on January 1, 2019)
21. A. Geiger, P. Lenz, and R. Urtasun, "Are We Ready for Autonomous Driving the Kitti Vision Benchmark Suite," in *Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp. 3354-3361, 2012
22. "Udacity. Public Driving Dataset," (https://www.udacity.com/self-driving-car, last accessed on January 10, 2019)
23. O. Russakovsky, J. Deng, and H. Su, "Imagenet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, Vol. 115, No. 3, pp. 211-252, 2015

Jianmin Liu received his B.S. degree from the University of Hunan, his M.S. degree from the University of Xiamen, and his Ph.D. from Central South University. His research interests include data mining and machine learning.

Minhua Yang received his Ph.D. from China Agricultural University. He is currently a professor at Central South University.

Jianmei Tan received her M.S. degree from the University of Technology of Changsha.