ILLUMINATING VEHICLES WITH MOTION PRIORS FOR SURVEILLANCE VEHICLE DETECTION

Xiaolian Wang[1,2], *Xiyuan Hu*[3], *Chen Chen*[1,2(✉)], *Zhenfeng Fan*[1,2], *Silong Peng*[1,2,4]

[1]Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Nanjing University of Science and Technology, Nanjing, China
[4]Beijing ViSystem Corporation Limited, Beijing, China

ABSTRACT

Vehicle detection in traffic surveillance videos is a special subtask in object detection, where desired objects are vehicles moving on the road while the background is still within a sequence. The disparity of speed within each frame, *i.e.* moving and static, is consistent with the vehicle and background semantic to some extent, thus motions can be extracted to enhance the appearance of foreground. In this paper, we propose a motion prior embedded parallel architecture for vehicle detection, aiming at illuminating vehicles and suppressing false positives in the background. We further implement extensive experiments on the UA-DETRAC dataset to validate the effectiveness of our approach, and achieve promising performance in both accuracy and speed.

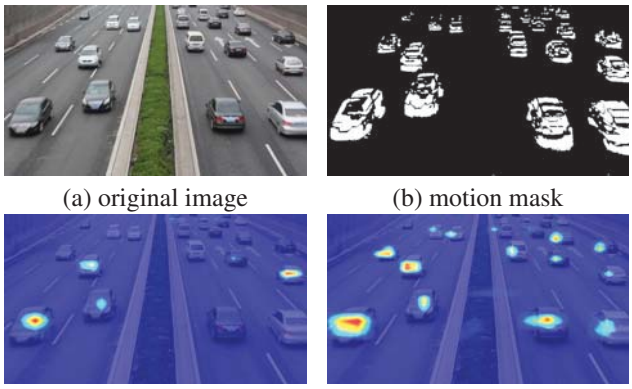
***Index Terms*—** Motion priors, vehicle detection, traffic surveillance videos

1. INTRODUCTION

Vehicle detection aims at simultaneously recognizing and localizing vehicles in images or frames. This is a crucial application in traffic surveillance, since accurate vehicle detection benefits downstream missions like vehicle tracking and re-identification. It also remains a challenging issue due to unconstrained environment such as lighting and occlusions, which imposes dramatic impacts on object appearance and raises many difficulties for detecting vehicles in real traffics.

Considering that the background in a traffic surveillance video is identical and only vehicles are moving along the road, we propose a parallel architecture embedded with motion priors to improve vehicle detection. The motivation stems from the sensitivity of human eyes to moving objects. The motion stimulation along with specific appearance can together highlight vehicles from the cluttered background.

Instead of simply blending motion priors into the detector, we decouple the moving object detection from the vehicle detection task by constructing a network of two partly sharing weights sub-branches. One branch takes both original frames and motion masks as input to separately detect moving vehicles. The other branch detects overall vehicles to ensure a full coverage of targets. The moving and global (static and moving) responses of vehicles are further aggregated for the final inference. By decoupling detection targets, motion priors can directly act on moving vehicles, and thus a false suppression from motion priors on static vehicles waiting for traffic lights can be avoided. The network structure will be detailed in Section 3. We further carry out experiments on the UA-DETRAC [1] dataset to demonstrate the effectiveness of our detector.



(a) original image (b) motion mask (c) baseline detector (d) our detector

Fig. 1. Effect of motion priors. Heatmaps (c-d) generated by Grad-CAM highlight the attention of detectors in identifying vehicles. The more and stronger responses triggered in (d) indicate our detector better focuses on vehicles when integrated with motion priors.

2. RELATED WORK

Generic object detection. Recent years have witnessed significant progress in object detection. As a classic type, two-stage detectors such as Faster R-CNN [2] first generate a set of candidate proposals, and then refine them for accurate bounding boxes and class labels. R-FCN [3] further intro-

978-1-7281-6395-6/20/$31.00 ©2020 IEEE 2021 ICIP 2020

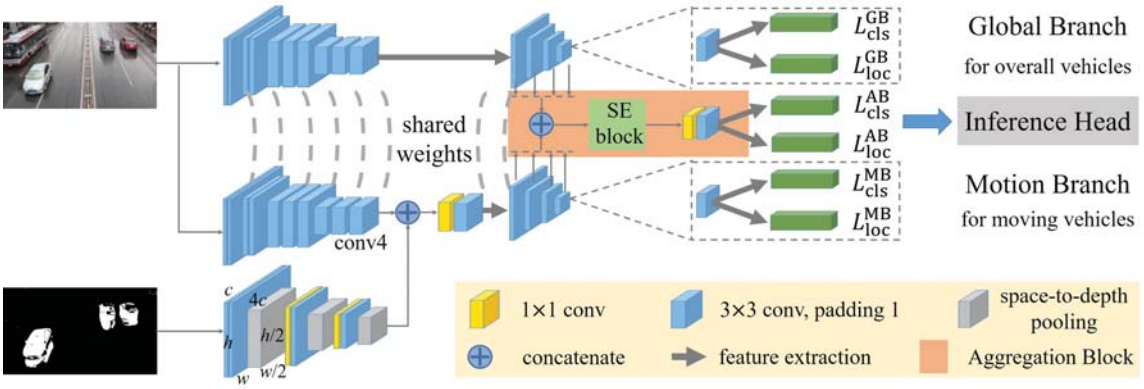


Fig. 2. An overview of the proposed framework. Here we show one prediction layer of feature pyramids for clarity.

duces position-sensitive score maps to improve the efficiency of Faster R-CNN. As for single-stage approaches which are popularized by SSD [4] and YOLO [5], objects are detected within a single network. Following that, many works focus on improving detection performance while maintaining computational efficiency such as RefineDet [6] and RetinaNet [7]. These image-based detectors explore representative features in geometric space, and thus can serve as the basic framework in our method for integrating motion priors.

Surveillance vehicle detection. In recent literature, EB [8] proposes a cascade detector called Evolving Boxes to refine bounding boxes by combining features with different fusion techniques. GP-FRCNNm [9] takes structural information of scenes into account and re-ranks proposals with an approximate geometric estimation of roads. FG-BRNet [10] separates foreground and background to generate gating features for suppressing false positives, and requires a feedback connection from detection results to distill foreground objects. These methods achieve remarkable performance to facilitate the application of vehicle detection in the surveillance system.

3. PROPOSED METHOD

The overview of our framework is illustrated in Fig. 2. Our detector contains three core modules, *i.e.* Motion Branch, Global Branch and Aggregation Block. Both branches are constructed on the same meta-detector like RefineDet [6] without loss of generality. In this section, we present details of our detector and explain the main purpose of design.

3.1. Motion priors embedding

Motions are commonly encoded as optical flow [11, 12] in video object detection [13, 14, 15]. Considering the characteristic of surveillance videos, we prefer a simpler and faster method, background subtraction such as Visual Background Extractor (ViBE) [16], to extract binary motion priors from past frames. We take the motion mask as input and integrate it into Motion Branch through a shallow side branch as shown in Fig. 2. Here we apply space-to-depth pooling [17] for downsampling feature maps, which is an operation stacking adjacent features of high resolution maps into channels, to retain complete motions as no pixels are discarded. Downsized motion masks are then concatenated with the `conv4_3` layer in VGG-16 [18], the backbone of RefineDet, to enrich semantics of low level layers. We also experiment on more complex structures but the gain is limited. Thus we choose this effective design for a fast processing speed despite its simpleness.

3.2. Parallel branches with shared weights

Decoupling moving objects from vehicle detection is an essential part in our method. Since static vehicles like those waiting for traffic lights or buses waiting for passengers are deactivated on motion masks, their appearance may be falsely weakened or eliminated as background when directly fusing features with motion priors. Thus we propose a parallel structure to separately detect moving vehicles to mitigate this negative effect. Global Branch is designed to be parallel with Motion Branch. It detects overall vehicles to ensure a full coverage of targets besides taking charge of all static foreground.

We further share weights between identical layers of Global and Motion Branch to narrow parameter space. This design is also based on the observation in experiments that sharing weights benefits accuracy of detectors. One possible reason may be that sharing weights enables Global Branch with no explicit motion attention to sense features highlighted in Motion Branch, while in turn, Global Branch stabilizes the training of vehicles in Motion Branch which lacks explicit supervision for static objects. Thus two branches together push features towards a compact space for vehicle detection.

3.3. Aggregation of moving and overall detection

Detection results of Motion and Global Branch need to be aggregated in inference, thus we employ Aggregation Block to

Table 1. Ablation studies on the UA-DETRAC *val* set.

input size	320×320			
baseline	√	√	√	√
+ Motion Branch		√	√	√
+ Aggregation Block			√	√
with shared weights		√		√
AP(%)	72.8	77.0	76.5	77.5
input size	512×512			
baseline	√	√	√	√
+ Motion Branch		√	√	√
+ Aggregation Block			√	√
with shared weights		√		√
AP(%)	75.0	79.8	78.9	80.7

get final vehicle predictions. Specifically, prediction layers of two branches are first concatenated with each other. We then insert a Squeeze-and-Excitation [19] (SE) block to generate channel-wise weights to recalibrate the concatenated features, aiming at adaptively enhancing informative patterns from different branches. Finally two convolutional layers are attached to the SE block to reduce feature dimensions as well as to extract features, on which bounding boxes can further be classified and regressed as those in Global and Motion Branch.

3.4. Joint training of multi-modules

As for the training of the whole network, we follow the loss function of classification (cls) and localization (loc) in [6] and jointly optimize the three modules as below:

$$L = \sum_{i \in \{\text{MB,GB,AB}\}} (L^i_{\text{cls}} + \alpha L^i_{\text{loc}}), \tag{1}$$

where *MB*, *GB* and *AB* are short for *Motion Branch*, *Global Branch* and *Aggregation Block*, respectively, and α is a weight term we set to 1. It is worth mentioning that the three modules are all activated in training, but we obtain final predictions only from Aggregation Block in inference.

Different from Global Branch and Aggregation Block which apply overall ground-truth boxes for supervision, Motion Branch is only optimized with *moving* labels. Since the separate moving labels for vehicles are not available in the UA-DETRAC dataset, we specify them according to the area of motion priors. We consider ground-truth boxes whose overlap with motion masks is higher than a threshold μ to be moving targets. This threshold should be set a bit higher than 0 to cope with noisy masks introduced by ViBE. Others such as static and slowly moving ones whose overlap with motion masks is less than μ are ignored as neutrals rather than negative background. By separately training moving vehicle detection, motion priors can directly act on moving objects and be fully fused into their appearance features without affecting the performance on static vehicles.

4. EXPERIMENTS

4.1. Experimental details

We conduct experiments on the challenging UA-DETRAC [1] dataset to validate the effectiveness of our method. This dataset contains 100 video sequences (60 for training and validation, 40 for testing) corresponding to more than 140,000 frames of real-world traffic scenes, and involves four scenarios of weather conditions, *i.e.* cloudy, rainy, sunny and night.

In the experiments, we set RefineDet as the baseline and optimize all variant detectors with Adam [22]. The learning rate is 10^{-4} for the first 32 epochs, then reduced to 10^{-5} and 10^{-6} for another two 8 epochs. We set the batch size to 10 for models with the input size of 320×320, and 6 for 512×512 models. The motion priors are generated by ViBE with default settings, which takes a negligible average time of 1.0ms per image on GPU. The threshold μ which defines moving labels is set to 0.1. According to [1], results evaluated by Average Precision (AP) at the matching IoU of 0.7 are reported.

4.2. Ablation studies

Our proposed detector mainly involves three components, *i.e.* Motion Branch with motion priors embedded, Aggregation Block, and shared weights within the parallel networks. We implement ablation studies to evaluate these components.

We first add Motion Branch and Aggregation Block to the baseline in sequence, and report results in Table 1. Since no aggregation is available in the + *Motion Branch* setting, we gather detections from Motion and Global Branch, and operate the common post-processing of Non-Maximum Suppression (NMS) afterwards to obtain final predictions. In the 320×320 model, both modules improve the detection performance. Applying Motion Branch contributes +4.2% (77.0%) AP, and learning to aggregate with a light neural network further improves the detector by 0.5% (77.5%) AP. Similar improvements are observed in the 512×512 model, where we achieve +4.8% and +0.9% AP gains for successively adding Motion Branch and Aggregation Block, respectively.

To validate the effectiveness of sharing weights within the parallel networks, we build a variant model with no weights shared between Motion Branch and Global Branch. As shown in Table 1, removing shared weights leads to 1.0% (76.5%) and 1.8% (78.9%) AP drop for detectors with the input size of 320×320 and 512×512, respectively. These results show that sharing weights can promote the feature learning for better vehicle detection even with half less parameters.

4.3. Visualized results

We further turn to Grad-CAM [23], a technique of producing visual explanations for decisions of Convolutional Neural Network (CNN)-based models, to see how our detector works. As shown in Fig. 1, the highlighted areas represent the

Fig. 3. Qualitative results of the baseline (top) and ours (bottom). As shown in (a), our detector suppresses the parterre that is falsely recognized by the baseline detector for its streamline shape similar with cars, and detects small vehicles with high confidence. It also steadily localizes the bus and truck in (b), and filters out the falsely detected traffic sign under a low illumination condition. Here yellow boxes represent false positives, and yellow arrows point to missed detections.

Table 2. Comparisons with the state-of-the-arts on the UA-DETRAC *test* set alongside the best/second best AP (%) results.

Method	Overall	Easy	Medium	Hard	Cloudy	Night	Rainy	Sunny	FPS	Environment
YOLOv2 [17]	57.72	83.28	62.25	42.44	57.97	64.53	47.84	69.75	-	GPU@GTX1080
Faster R-CNN [2]	58.45	82.75	63.05	44.25	66.29	69.85	45.16	62.34	11.1	GPU@TitanX
EB [8]	67.96	89.65	73.12	53.64	72.42	73.93	53.40	83.73	10	GPU@TitanX
R-FCN [3]	69.87	93.32	75.67	54.31	74.38	75.09	56.21	84.08	6	GPU@TitanX
CSP [20]	77.67	93.65	83.67	64.54	86.81	80.63	61.39	89.66	4	GPU@K40
GP-FRCNNm [9]	77.96	92.74	82.39	67.22	83.23	77.75	70.17	86.56	4	GPU@K40
HAT [21]	78.64	93.44	83.09	68.04	86.27	78.00	67.97	88.78	3.6	GPU@TitanX
FG-BR Net [10]	79.96	93.49	83.60	70.78	87.36	78.42	70.50	89.89	10	GPU@M40
Ours	80.76	94.56	85.90	69.72	87.19	80.68	71.06	89.74	14	GPU@GTX1080

attention of detectors in identifying vehicles. Our detector perceives a broader range of vehicles in correspondence with motion masks compared with the baseline detector.

The qualitative performance is shown in Fig. 3 with a score threshold of 0.5 for displaying. Since the background in surveillance videos hardly changes, false positives such as the parterre and traffic signs which have similar shapes with cars may continuously appear once detected as shown in the top line of Fig. 3. Compared with the baseline detector, our detector performs more stably in recalling true positives as well as suppressing false positives in real traffic background.

4.4. Comparisons with the state-of-the-arts

Following the protocol of UA-DETRAC, we submit the results of our detector with the input size of 512×512 to the public testing server for evaluation. A comparison with recently published state-of-the-art methods is shown in Table 2. We achieve an overall accuracy of 80.76% AP while maintaining the fastest speed of 14 FPS among these detectors. In terms of the performance under different weather conditions, our approach obtains competitive results on the *cloudy* and *sunny* subset, and outperforms the other methods on the *night* and *rainy* subset. We attribute the stable performance under various conditions especially the bad weather to the proper use of motion priors. Detectors only use geometric features are susceptible to unexpected environment when detecting vehicles in real traffic, thus motions are very critical to generate robust predictions in surveillance vehicle detection.

5. CONCLUSION

In this paper, we propose the motion priors embedded parallel architecture for surveillance vehicle detection. The key is to properly leverage motions by decoupling moving objects from overall vehicles, in order to enhance vehicle appearance while carefully suppressing false positives in the background.

Acknowledgments.

The corresponding author is Chen Chen (chen.chen@ia.ac.cn). This work was supported by the National Science Foundation of China under Grant NSFC 61906194, and National Key Research and Development Project 2018YFC0807306.

6. REFERENCES

[1] L. Wen, D. Du, Z. Cai, Z. Lei, M. C. Chang, H. Qi, J. Lim, M. H. Yang, and S. Lyu, "DETRAC: A new benchmark and protocol for multi-object detection and tracking," *arXiv preprint arXiv:1511.04136*, 2015.

[2] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions of Pattern Analysis and Machine Intelligence, TPAMI*, vol. 39, no. 6, pp. 1137–1149, 2017.

[3] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Conference and Workshop on Neural Information Processing Systems, NeurIPS*, 2016.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.Y. Fu, and A.C. Berg, "SSD: Single shot multibox detector," in *European conference on computer vision, ECCV*, 2016, pp. 21–37.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2016, pp. 779–788.

[6] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2018, pp. 4203–4212.

[7] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollr, "Focal loss for dense object detection," in *IEEE International Conference on Computer Vision, ICCV*, 2017.

[8] L. Wang, Y. Lu, H. Wang, Y. Zheng, H. Ye, and X. Xue, "Evolving boxes for fast vehicle detection," in *IEEE International Conference on Multimedia and Expo, ICME*, 2017, pp. 1135–1140.

[9] S. Amin and F. Galasso, "Geometric proposals for faster r-cnn," in *IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS*, 2017, pp. 1–6.

[10] Z. Fu, Y. Chen, H. Yong, R. Jiang, L. Zhang, and X. S. Hua, "Foreground gating and background refining network for surveillance object detection," *IEEE Trans. Image Processing*, vol. 28, no. 12, pp. 6077–6090, 2019.

[11] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[12] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazirbas, V. Golkov, P. Smagt, D. Cremers, and T. Brox, "FlowNet: Learning optical flow with convolutional networks," in *IEEE International Conference on Computer Vision, ICCV*, 2015, pp. 2758–2766.

[13] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei, "Deep feature flow for video recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2017, pp. 4141–4150.

[14] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei, "Flow-guided feature aggregation for video object detection," in *IEEE International Conference on Computer Vision, ICCV*, 2017, pp. 408–417.

[15] X. Zhu, J. Dai, and L. Yuan, "Towards high performance video object detection," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2018, pp. 7210–7218.

[16] O. Barnich and M. V. Droogenbroeck, "ViBE: A powerful random technique to estimate the background in video sequences," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2009, pp. 945–948.

[17] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2017, pp. 6517–6525.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR*, 2015.

[19] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2018, pp. 7132–7141.

[20] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2019.

[21] S. Wu, M. Kan, S. Shan, and X. Chen, "Hierarchical attention for part-aware face detection," *International Journal of Computer Vision, IJCV*, vol. 127, no. 6-7, pp. 560–578, 2019.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR*, 2015.

[23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *IEEE International Conference on Computer Vision, ICCV*, 2017, pp. 618–626.