ACCURATE VEHICLE DETECTION USING MULTI-CAMERA DATA FUSION AND MACHINE LEARNING

Hao Wu, Xinxiang Zhang, Brett Story, Dinesh Rajan

Department of Electrical Engineering, Southern Methodist University, Dallas, TX 75025, USA

ABSTRACT

Computer-vision methods have been extensively used in intelligent transportation systems for vehicle detection. However, the detection of severely occluded or partially observed vehicles due to the limited camera fields of view remains a challenge. This paper presents a multi-camera vehicle detection system that significantly improves the detection performance under occlusion conditions. The key elements of the proposed method include a novel multi-view region proposal network that localizes the candidate vehicles on the ground plane. We also infer the vehicle position on the ground plane by leveraging multi-view cross-camera context. Experiments are conducted on dataset captured from a roadway in Richardson, TX, USA, and the system attains 0.7849 Average Precision and 0.7089 Multi Object Detection Precision. The proposed system results in an approximately 31.2% increase in AP and 8.6% in MODP than the single-camera methods.

Index Terms—Vehicle detection, vehicle occlusion, multi-camera, region proposal network, multi-view fusion.

1. INTRODUCTION

Vision-based vehicle detection methods have recently received significant attention in intelligent transportation systems (ITSs). Reliable vehicle detection is a fundamental component of traffic surveillance with increased safety and mobility implications [1]. A comprehensive review of vehicle detection system is given in [2]. With the recent resurgence of deep learning algorithms in a wide range of fields including image processing and pattern recognition [3-5], convolutional neural network (CNN)-based single-camera object detection systems have been studied [6-10]. However, these single-camera systems are not able to detect partially-occluded vehicles in crowded traffic scene. One way to overcome the challenge of detecting partially-occluded vehicles is to detect the candidate vehicles using their multiple semantic sub-parts [13-15]. Although these methods adapt to situations with partial occlusions, they fail when vehicles are severely occluded in traffic dynamics [11, 12, 17]. Another approach to overcome the occlusion challenge is to use a multi-camera system and fuse the information from each independent camera stream [16, 20]. Recent algorithms on multi-camera object detection mainly focus on pedestrian detection. These algorithms infer the pedestrian locations on the ground plane by extracting monocular features and estimating the ground-plane occupancy vector. In order to estimate the ground-plane occupancy vector, some of the multi-camera object detection systems extract binary foreground mask as the feature, which is not robust in severely-occluded traffic scenes [18, 19, 21]. Some other algorithms use features generated by deep CNN [22, 23]. The existing approaches fuse the extracted features to infer the occupancy vector. The location of a pedestrian is represented by a single ground-plane cell with predefined shape and size [18, 19, 21-23]. The fixed-size cells are appropriate for detecting pedestrians due to the similarity of the footprint of various pedestrians on ground plane. However, using fixed cells to detect vehicles that have large variations in shape and size, e.g. truck vs. sedan on the ground plane is not appropriate. Moreover, in [22], the occupancy vector on the ground plane is obtained from each side view independently; and in [23], the estimation of the multi-view joint occupancy requires higher computational complexity to project every ground-plane cell back to each side view.

Therefore, to address the aforementioned issues, this paper develops: 1) a Multi-View Region Proposal Network (MVRPN) to estimate the ground-plane occupancy vector by leveraging multiple side views simultaneously, and 2) a fine-tuned pre-trained deep CNN to remove false positive object predictions that are generated by the trained MVRPN. In the proposed system, the MVRPN is trained by using given ground-plane information, which is captured from a top-view camera. Instead of using cells with a single predefined size, the location of objects on the ground plane are represented by cell blocks with adaptive size. Therefore, the proposed system can be applied to vehicles with large variations in size.

The remainder of this paper is organized as follows. Section 2 presents the descriptions of the proposed multi-view vehicle detection system. Experiments and results are provided in Section 3, followed by conclusions in Section 4.

2. PROPOSED SYSTEM DESCRIPTION

The core objective of the proposed system is to localize the vehicles on the ground plane by fusing synchronized frames from a multi-camera network. An overview of the proposed system is shown in Fig. 1, and the frequently used notations are given in Table 1. A MVRPN is introduced to deduce the candidate vehicle Region of Interests (ROIs) on the ground

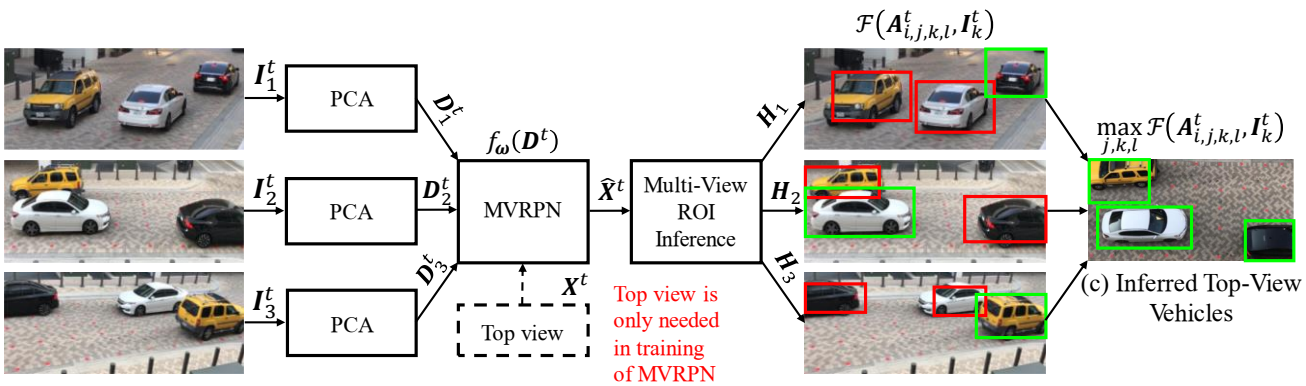


Fig. 1. The overview of the multi-camera vehicle detection system. The original synchronized frames from 3 side cameras are shown in (a). The detected vehicles on side views and the inferred vehicles on top view are shown in (b) and (c). The top-view vehicles are inferred by the corresponding detections with the maximum probabilities, which are the green boxes in (b).

Table 1. Frequently Used Notations

	Description
$\boldsymbol{I}_k^t$	The t^{th} RGB frame from side view k.
$\boldsymbol{G}^t$	The 2-D binary ground-plane grid of cells.
$\boldsymbol{X}^t$	The 1-D ground-truth Boolean occupancy vector.
$\boldsymbol{D}^t$	The dimension-reduced input vector.
$\widehat{\boldsymbol{X}}^t$	The estimated ground-plane occupancy vector.
$\boldsymbol{R}_i^t$	The i^{th} MER on the ground plane.
$\boldsymbol{H}_k$	The homography between k^{th} side view and ground plane.
$\boldsymbol{C}_{i,j}^t$	The j^{th} foreground cells in $\boldsymbol{R}_i^t$ on the ground plane.
$\boldsymbol{P}_{i,j,k}^t$	The projection of top-left corner $\boldsymbol{C}_{i,j}^t$ in $\boldsymbol{R}_i^t$ at side view k.
$\boldsymbol{A}_{i,j,k,l}^t$	The l^{th} bounding box with bottom edge centered at $\boldsymbol{P}_{i,j,k}^t$.
$f_{\boldsymbol{\omega}}(\cdot)$	The function that represents the MVRPN.
$\mathcal{F}(\cdot)$	The fine-tuned pre-trained deep CNN classification.

*Note: $\boldsymbol{R}_i^t$, $\boldsymbol{C}_{i,j}^t$ and $\boldsymbol{A}_{i,j,k,l}^t$ are 4-element vectors that represent the selected rectangular bounding boxes with the form $[x_{\min}\ y_{\min}\ \omega\ h]$.

plane from side-view images. A multi-view ROI inference is then used to obtain the probability of the deduced ROIs being a vehicle.

Consider a camera network composed of C side-view cameras and 1 top-view camera, where each camera can have different resolution. The top-view camera is used to capture the ground-truth information from ground plane with less occlusion for MVRPN and to quantify the performance of the proposed algorithm; a top-view camera is not necessary for field implementation of a trained system. The t^{th} RGB frame captured from side-view camera k is denoted as $\boldsymbol{I}_k^t$ with size equal to $N_k \times M_k \times 3$, where $k \in \{1,2,\dots,C\}$. The large dimension of the input $\boldsymbol{I}_k^t$ increases the unknown training parameters and makes the MVRPN computationally hard to converge [24]. Hence Principle Component Analysis (PCA) [25] is used to generate $\boldsymbol{D}_k^t$, a n_k dimensional column vector from $\boldsymbol{I}_k^t$, where $n_k \ll N_k \cdot M_k \cdot 3$. In this study, we set $n_k = 500$. From the top-view camera, the t^{th} $N_G \times M_G \times 3$ ground-plane frame is captured. A foreground binary mask is then obtained by binary pixel-wise labeling of the ground-plane frame into the vehicle and non-vehicle class. The binary mask of the ground-plane frame is subsampled into a $\frac{N_G}{m} \times \frac{M_G}{m}$ grid of cells, where m is a hyper parameter to adjust the size of grid of cells while ensuring its aspect ratio is identical to the ground-plane frame. We set $m = 20$, and the total number of cells is $N = \frac{N_G}{m} \times \frac{M_G}{m}$. We denote the grid of cells as a 2-D binary matrix G^t, where the matrix element with value equal to 1 represents the corresponding cell is occupied by a vehicle. By concatenating columns of G^t, the $N \times 1$ ground-truth Boolean occupancy vector is obtained. We denote occupancy vector as $\boldsymbol{X}^t$, where $\boldsymbol{X}^t = \{X_1^t, X_2^t, \dots X_N^t\}^T$. Note that the superscript t of those notations is the index into the set of captured frames.

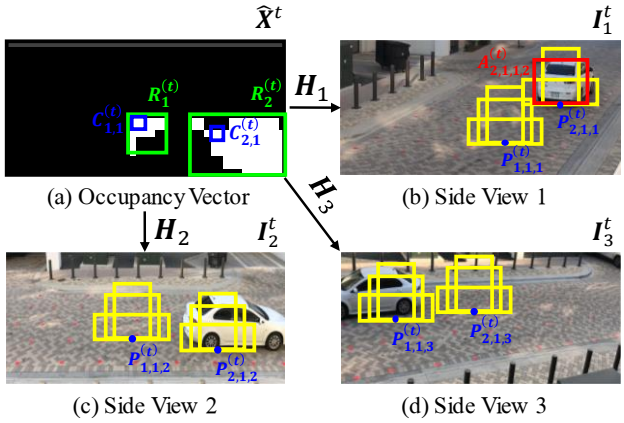
2.1. Multi-View Region Proposal Network (MVRPN)

After the PCA procedure, the input column vector $\boldsymbol{D}^t = \{\boldsymbol{D}_1^t, \boldsymbol{D}_2^t, \dots, \boldsymbol{D}_c^t\}^T$ of the MVRPN is obtained, where $\boldsymbol{D}^t$ is composed of C dimension-reduced vector of frames captured from different side-view cameras at the same time. Given $\boldsymbol{D}^t$, a Multi-Layer Perceptron (MLP) architecture, MVRPN, is utilized to estimate the ground-plane occupancy vector, $\widehat{\boldsymbol{X}}^t = \{\hat{X}_1^t, \hat{X}_2^t, \dots, \hat{X}_N^t\}^T$. In the proposed system, we assume that the number of cells occupied by vehicle on the ground plane is less than those corresponding to background. Therefore, due to the imbalanced vehicle instances, the training process of MVRPN suffers from the bias problem [26]. To alleviate this issue, the loss function $\boldsymbol{\mathcal{L}}$ in training the MVRPN is set as:

$$\boldsymbol{\mathcal{L}}_{\boldsymbol{\omega}}(\boldsymbol{X}^t, \widehat{\boldsymbol{X}}^t) = \begin{cases} \frac{\alpha}{2N} \sum_{i=1}^{N} \left[\hat{X}_i^t - X_i^t\right]^2, \text{if } X_i^t = 1 \\ \frac{1}{2N} \sum_{i=1}^{N} \left[\hat{X}_i^t - X_i^t\right]^2, \text{otherwise} \end{cases} \quad (1)$$

where

$$\widehat{\boldsymbol{X}}^t = f_{\boldsymbol{\omega}}(\boldsymbol{D}^t) = \{\hat{X}_1^t, \hat{X}_2^t, \dots, \hat{X}_N^t\}^T \quad (2)$$



(a) Occupancy Vector (b) Side View 1

(c) Side View 2 (d) Side View 3

Fig. 2. The probability assignment process. The ground-plane locations of MER and its foreground cell are shown in (a). The corresponding projection of top-left corner of foreground cell is shown at each side view. The red bounding box is assigned the maximum cell probability across all side views.

The loss function $\mathcal{L}_{\omega}(\boldsymbol{X}^t, \hat{\boldsymbol{X}}^t)$ is the weighted Mean Squared Error (WMSE) between the estimated $\hat{\boldsymbol{X}}^t$ and the ground truth $\boldsymbol{X}^t$. MVRPN is denoted as $f_{\omega}(\boldsymbol{D}^t)$, where $\boldsymbol{\omega}$ are the MVRPN parameters to be learned, and $\hat{\boldsymbol{X}}^t$ is the output of MVRPN. The penalization weight α adaptively applies more penalty to the computed WMSE when MVRPN classifies a foreground cell as background. In this study, we set $\alpha = 5$.

2.2. Multi-View ROI Inference

After estimating the occupancy vector $\hat{\boldsymbol{X}}^t$, a set of candidate ROIs, which are Minimum Enclosing Rectangles (MERs) to enclose foreground cells block, are generated. In this paper, the i^{th} MER of $\hat{\boldsymbol{X}}^t$ is denoted as $\boldsymbol{R}_i^t$, where $i \in \{1, 2, \dots, M\}$, and M is the total number of MERs on the ground plane. The j^{th} foreground cells in i^{th} MER is denoted as $\boldsymbol{C}_{i,j}^t$, where $j \in \{1, 2, \dots, P\}$ and P is the number of foreground cells within $\boldsymbol{R}_i^t$. However, since some MERs are false positives (FPs), a multi-view ROI inference is leveraged to remove those FPs. For this purpose, a set of homography matrices are estimated by using RANSAC and Levenberg-Marquardt algorithms [27], where a set of red markers shown in Fig. 2 are used to generate the corresponding interest points. The homography matrix that represents the relationship between k^{th} side view and the ground plane is denoted as $\boldsymbol{H}_k$, where $k \in \{1,2, \dots, C\}$ and C is the number of side-view cameras. From the estimated homography matrices, the top-left corner of $\boldsymbol{C}_{i,j}^t$ in every MER is projected to each side view. We denote the projected pixel of the top-left corner of $\boldsymbol{C}_{i,j}^t$ in side view k as $\boldsymbol{P}_{i,j,k}^t$. A set of bounding boxes are then generated according to the projected pixels, where each pixel associates with L multi-scale-multi-aspect-ratio bounding boxes. The bottom edge of each bounding box is centered at the corresponding projected pixel [18]. We denote the l^{th} bounding box whose bottom edge centered at the $\boldsymbol{P}_{i,j,k}^t$ as $\boldsymbol{A}_{i,j,k,l}^t$, where $l \in \{1,2, \dots, L\}$ and L is the total number of bounding boxes associated with $\boldsymbol{P}_{i,j,k}^t$. In this work, bounding boxes with 3 different scales and 3 different aspect ratios are used, and hence $L = 9$ for each projected pixel. AlexNet, a pre-trained deep CNN $\mathcal{F}(\boldsymbol{A}_{i,j,k,l}^t, \boldsymbol{I}_k^t)$ is fine-tuned by transfer learning to assign the probability to each bounding box. The maximum probability of the bounding box being vehicle is assigned to the MER $\boldsymbol{R}_i^t$ on the ground plane as:

$$\Pr(\boldsymbol{R}_i^t | G^t) = \max_{j,k,l} \mathcal{F}(\boldsymbol{A}_{i,j,k,l}^t, \boldsymbol{I}_k^t) \tag{3}$$

The probability assignment process is illustrated in Fig. 2, where $\boldsymbol{A}_{2,1,1,2}^t$ is assigned the maximum probability and the false positive MER $\boldsymbol{R}_1^t$ is eliminated by multi-view ROI inference. The state $S(\boldsymbol{R}_i^t | G^t)$ of the MER $\boldsymbol{R}_i^t$ is estimated using probability thresholding as:

$$S(\boldsymbol{R}_i^t | G^t) = \begin{cases} 0, & \text{if } \Pr(\boldsymbol{R}_i^t | G^t) \le a \\ 1, & \text{otherwise} \end{cases} \tag{4}$$

where $a \in [0,1]$ is the probability threshold. The threshold a is determined such that the prediction results yield the highest performance in validation set. The proposed system recalls $\boldsymbol{R}_i^t$ as the vehicle when $S(\boldsymbol{R}_i^t | G^t) = 1$ and eliminates $\boldsymbol{R}_i^t$ when $S(\boldsymbol{R}_i^t | G^t) = 0$.

3. EXPERIMENTS

In this section, we present experimental results of the proposed automatic multi-camera vehicle detection system. The experiments are conducted on real-traffic image data that is captured from a roadway at Richardson, TX, USA.

3.1. Data Preparation

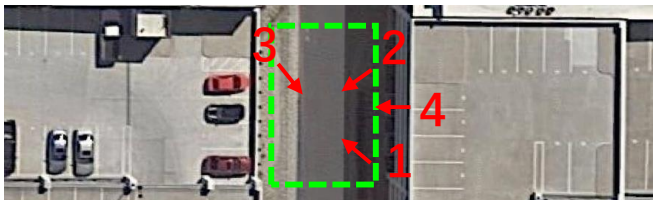


Fig. 3. The multi-camera network setup. Note: location 1, 2, 3 are side-view cameras, and location 4 is top-view camera.

The synchronized image data is captured from 4 cameras as shown in Fig. 3. The captured frames are sampled such that the number of frames with vehicles are equal to those without vehicles. The remaining 9960×4 frames are split into the ratio 3:1:1 correspondingly to training, validation and test sets. For MVRPN training, the synchronized dimension-reduced frames of 3 side cameras are used as inputs. The target top-view frames are labeled as pixel-wise binary masks, where the positive values (+1) indicate the vehicle and the negatives (0) indicate the background on the ground plane. Note that

the training samples are input into the MVRPN randomly rather than chronologically. For CNN training, the ground-truth bounding boxes are labeled at 3 side views, and image patches are then extracted by applying Edge Boxes [28]. The extracted image patches whose Intersection over Union (IoU) with a ground-truth bounding box greater than 0.7 are treated as positives; IoU less than 0.3 are treated as negatives; and the rest are ignored. The ratio of the positive samples to the negative samples is set to 1:2.

3.2. Model Training Configuration

All the experiments are performed using a desktop with Intel (R) Quad-Core (TM) i5-7400 CPU@3.0GHz Processor, 8GB RAM, and NVIDIA GeForce GTX 1050Ti 4GB GPU.

3.2.1. Multi-view region proposal network

The MVRPN is trained by minimizing the loss function in Eq. 1. The synchronized side-view frames are RGB images. The 1500×1 MVRPN input vector is obtained by retaining the first 500 principal components for each of the 3 side views. Ground-truth occupancy vectors are obtained by subsampling 300×600 ground-plane binary mask into 15×30 grid of cells. During training process, RMSProp [29] with 128 batch size, 0.15 initial learning rate, $\eta^+ = 1.2$, and $\eta^- = 0.5$ is applied.

3.2.2. Transfer learning prediction

The fine tuning of the pre-trained AlexNet is implemented on MATLAB R2017b with AlexNet support package. During the training process, stochastic gradient descent (SGD) [30] with 128 batch size, 0.9 momentum, 10^{-4} initial learning rate, and 10^{-4} L_2 regularization is applied.

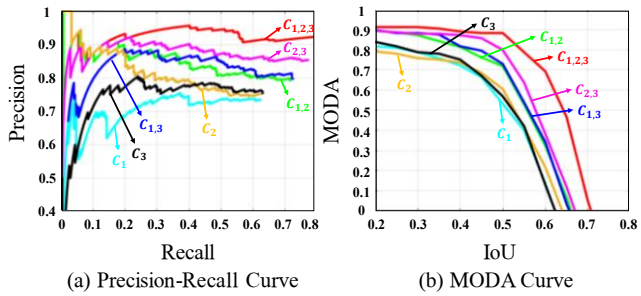
3.3. Comparative Evaluation

Table 2. Numeric Evaluation Results

Camera deployments	AP	MODP
$\boldsymbol{C_{1,2,3}}$	**0.7849**	**0.7089**
$C_{1,2}$	0.6087	0.6526
$C_{1,3}$	0.5989	0.6554
$C_{2,3}$	0.6761	0.6722
C_1	0.4401	0.6175
C_2	0.5124	0.6287
C_3	0.4673	0.6208

*Note: $C_{\alpha,\beta,\gamma}$ represents utilization of side camera α, β, and γ.

We evaluate the multi-camera vehicle detection system on 1992 top-view test images. To our best knowledge, there is no published dataset about multi-camera vehicle detection. The feature extracted in the existing multi-camera pedestrian detection algorithm is not applicable in this paper [17, 21-23]. Hence, we benchmark the performance of the multi-camera vehicle detection system by deploying different camera combinations. For the fixed IoU, the system is evaluated by Average Precision (AP) [31] and Multiple Object Detection Precision (MODP) [32]. The detected bounding boxes are



(a) Precision-Recall Curve (b) MODA Curve

Fig. 4. Evaluation curves. The Precision-Recall curve is shown in (a). The MODA curve is shown in (b).

considered as true positives when the IoUs exceed 0.55. The precision-recall curve is shown in Fig. 4(a). For the varying IoUs, Multiple Object Detection Accuracy (MODA) curve [32] is shown in Fig. 4(b). The evaluation results of AP and MODP are shown in Table 2, where the camera deployment $C_{1,2,3}$ achieves the best performance (0.7849 AP and 0.7089 MODP) among all variations. The utilizations of 2 side-view cameras achieve better performances than single camera deployments. Such numeric evaluation results indicate that the performance of the multi-camera vehicle detection system increases when more side-view cameras are deployed.

3.4. Visualization Results

(a) (b)

Fig. 5. Ground-plane detection results. Red bounding boxes are detections, and green bounding boxes are ground truths.

Examples of vehicles detected on the ground plane using $C_{1,2,3}$ are shown in Fig. 5. The system detects the vehicles with variant sizes, e.g. the white sedan vs. the yellow SUV in Fig. 5(a). The partially-observed black SUV with smaller size than regular vehicle is also detected in Fig. 5(a). However, the detected bounding box of the yellow vehicle at Fig. 5(a) is not of optimal shape and size, and the partially-observed vehicle at right boundary of Fig. 5(b) is not detected.

4. CONCLUSION

In this paper, a multi-camera vehicle detection system with a MVRPN/CNN pipeline is presented. The proposed system detects partially and severely occluded vehicles in field traffic scenes. In future investigations, a multi-view bounding-box regression will be embedded into the pipeline to optimize the bounding-box predictions. A vehicle detection system which can utilize temporal video frames will be developed to address vehicle tracking-related challenges. In addition, the optimal locations to place side cameras will also be studied.

5. REFERENCE

[1] L. E. Y. Mimbela, and L. A. Klein, "Summary of vehicle detection and surveillance technologies used in intelligent transportation systems," Federal Highway Administration, Tech. Rep., 2000.

[2] Z. Sun, G. Bebis, and R. Miller, "On-road vehicle detection: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 694-711, 2006.

[3] Y. Wang, Y. Huang, W. Zheng, Z. Zhou, D. Liu, and M. Lu, "Combining convolutional neural network and self-adaptive algorithm to defeat synthetic multi-digit text-based CAPTCHA," In *Proceedings of the IEEE International Conference on Industrial Technology*, Mar. 2017, pp. 980-985.

[4] Y. Wang, Z. Zhou, S. Jin, D. Liu, and M. Lu, "Comparisons and selections of features and classifiers for short text classification," In *IOP Conference Series: Materials Science and Engineering*, vol. 261, no. 1, pp. 012018, 2017.

[5] Y. Zhang, and X. Zhang, "Effective real-scenario Video Copy Detection," In *Proceedings of the IEEE International Conference on Pattern Recognition*, Dec. 2016, pp. 3951-3956.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "SSD: Single Shot Multibox Detector," In *Proceedings of the European Conference on Computer Vision*, Oct. 2016, pp. 21-37.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 779-788.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," In *Advances in Neural Information Processing Systems*, 2015, pp. 91-99.

[9] Y. Wang, and M. Lu, "An optimized system to solve text-based CAPTCHA," *International Journal of Artificial Intelligence and Applications*, vol.9, no.3, 2018.

[10] Y. Wang, and M. Lu, "A self-adaptive algorithm to defeat text-based CAPTCHA," In *Proceedings of the IEEE International Conference on Industrial Technology*, Mar. 2016, pp. 720-725.

[11] H. Possegger, T. Mauthner, P. M. Roth, and H. Bischof, "Occlusion geodesics for online multi-object tracking," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1306-1313.

[12] J. Chang, L. Wang, G. Meng, S. Xiang, and C. Pan, "Vision-Based Occlusion Handling and Vehicle Classification for Traffic Surveillance Systems," *IEEE Intelligent Transportation Systems Magazine*, vol. 10, no. 2, pp. 80-92, 2018.

[13] B. Tian, Y. Li, B. Li, and D. Wen, "Rear-view vehicle detection and tracking by combining multiple parts for complex urban surveillance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 2, pp. 597-606, 2014.

[14] B. Tian, M. Tang, and F. Y. Wang, "Vehicle detection grammars with partial occlusion handling for traffic surveillance," *Transportation Research Part C: Emerging Technologies*, vol. 56, pp. 80-93, 2015.

[15] C. Wang, Y. Fang, H. Zhao, C. Guo, S. Mita, and H. Zha, "Probabilistic inference for occluded and multiview on-road vehicle detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 1, pp. 215-229, 2016.

[16] M. Hofmann, D. Wolf, and G. Rigoll, "Hypergraphs for joint multi-view reconstruction and multi-object tracking," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3650-3657.

[17] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern recognition letters*, vol. 34, no. 1, pp. 3-19, 2013.

[18] K. Kim, and L. S. Davis, "Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering," In *Proceedings of European Conference on Computer Vision*, May. 2006, pp. 98-109.

[19] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 267-282, 2008.

[20] A. C. Sankaranarayanan, A. Veeraraghavan, and R. Chellappa, "Object detection, tracking and recognition for multiple smart cameras," *Proceedings of the IEEE*, vol. 96, no. 10, pp. 1606-1624, 2008.

[21] P. Peng, Y. Tian, Y. Wang, J. Li, and T. Huang, "Robust multiple cameras pedestrian detection with multi-view Bayesian network," *Pattern Recognition*, vol. 48, no. 5, pp. 1760-1772, 2015.

[22] P. Baqué, F. Fleuret, and P. Fua, "Deep occlusion reasoning for multi-camera multi-target detection," In *Proceedings of the IEEE International Conference on Computer Vision*, Oct. 2017, vol. 2.

[23] T. Chavdarova, "Deep multi-camera people detection," In *Proceedings of the IEEE International Conference on Machine Learning and Applications*, Dec. 2017, pp. 848-853.

[24] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1982-2001, 2010.

[25] I. Jolliffe, "Principal component analysis," In *International Encyclopedia of Statistical Science*, pp. 1094-1096, 2011.

[26] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431-3440.

[27] R. Hartley, and A. Zisserman, *Multiple view geometry in computer vision*, Cambridge university press, 2003.

[28] C. L. Zitnick, and P. Dollár, "Edge boxes: Locating object proposals from edges," In *Proceedings of the European Conference on Computer Vision*, Sep. 2014, pp. 391-405.

[29] T. Tieleman, and G. Hinton, "Lecture 6.5-RMSProp: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, pp. 26-31, 2012.

[30] L. Bottou, "Large-scale machine learning with stochastic gradient descent," In *Proceedings of COMPSTAT*, 2010, pp. 177-186.

[31] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303-338, 2010.

[32] R. Kasturi, et al, "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 319-336, 2009.