IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 29, 2017, accepted September 20, 2017, date of publication September 26, 2017,
date of current version November 14, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2756081

Scene-Adaptive Vehicle Detection Algorithm Based on a Composite Deep Structure

YINGFENG CAI[2], HAI WANG[1], ZHENGYANG ZHENG[1], AND XIAOQIANG SUN[2]
[1]Department of Automotive Engineering, School of Automotive and Traffic Engineering, Jiangsu University, Zhenjiang 212013, China
[2]Intelligent Vehicle Research Institute, Automotive Engineering Research Institute, Jiangsu University, Zhenjiang 212013, China

Corresponding author: Hai Wang (wanghai1019@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61601203, Grant 61403172, and Grant U1564201, in part by the Key Research and Development Program of Jiangsu Province under Grant BE2016149, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20140555, and in part by the Six Talent Peaks Project of Jiangsu Province under Grant 2015-JXQC-012 and Grant 2014-DZXX-040.

ABSTRACT Existing machine-learning-based vehicle detection algorithms for intelligent vehicles have an obvious disadvantage in that the detection effect decreases dramatically when the distribution of training samples and the scene target samples do not match. To address this issue, a scene-adaptive vehicle detection algorithm based on a composite deep structure is proposed in this paper. Inspired by the Bagging (Bootstrap aggregating) mechanism, multiple relatively independent source samples are first used to build multiple classifiers and then voting is used to generate target training samples with confidence scores. The automatic feature extraction ability of deep convolutional neural network is then used to perform source-target scene feature similarity calculations with a deep auto-encoder in order to design a composite deep-structure-based scene-adaptive classifier and its training method. Experiments on the KITTI data set and a data set captured by our group demonstrate that the proposed method performs better than existing machine-learning-based vehicle detection methods. In addition, compared with existing scene-adaptive object detection methods, our method improves the detection rate by an average of approximately 3%.

INDEX TERMS Image recognition, vehicle detection, scene adaptive, composite deep structure, deep convolutional neural network.

I. INTRODUCTION

Existing monocular-vision-based vehicle detection methods for intelligent vehicles generally fall into one of two categories: background-modeling-based methods and machine-learning-based approaches.

Background-modeling-based methods firstly capture multiple image frames to build a background model and then segment the foreground image by background subtraction. On this basis, the target belonging to vehicles can be further judged and extracted. This type of method applies only to a static surveillance scene with a fixed camera, and often fails for a dynamic background. In contrast, the machine-learning-based approach extracts features from the image and trains a vehicle classifier using a large number of training samples. The sub-images belonging to the vehicle targets are then verified in the road image with this classifier. Compared with background modeling, machine-learning-based methods are more robust to dynamic scenes and thus have gradually become the mainstream method used in current research. The core research activities of this approach relate to how the image features should be expressed and how the vehicle classifier should be built.

For feature expression, primary features with clear physical characteristics were often used in earlier studies, such as shadows, symmetry, texture and vertical edges [1], [2]. These primary features have limited expression ability and cannot describe the verities of vehicle classes. Therefore, many purpose-designed general image descriptors have been developed to improve the description ability of features, such as the Haar-like feature, HOG (histogram of orientated gradient), LBP (local binary patterns), SIFT (scale-invariant feature transform) and Wavelet feature [3].The classifier should be capable of accurately determining the optimal decision boundary, which is another factor that affects vehicle detection performance. The most popular classifiers used for the vehicle detection task are Adaboost and SVM (support vector machine). The former classifier is an adaptive Boosting algorithm which builds a strong classifier by combining

several weak classifiers with linear weights [4]. The latter classifier maps the data to a higher dimensional space and searches for the optimal classification surface, which is the surface with the largest interval between classes [5]. Recently, several improved methods based on these two classifiers have been proposed which further enhance the performance and speed of the object detection tasks [6], [7].

Although machine-learning-based vehicle detection performs well for many applications, some problems still exist that require solving. One of the most critical problems is the adaptability of the trained classifier. The existing training method usually collects and labels a large number of samples by hand to train a vehicle classifier. This classifier may work well for some scenes but may fail for other scenes when there is a large visual difference between the target scenes and the source scenes where the training samples came from, due to scene complexity, vehicle appearance and position differences. On the other hand, it is labor-intensive to recollect and relabeled samples from the target scene each time, and it also significantly reduces the degree of automation for vehicle detection tasks in a new target scene.

Traditional machine-based methods are based on statistical theory principles, which make the important assumption that the distribution between the source data and the target test data are the same. As the problem above shows, there are often large differences in the data distribution between the original training data set and the samples in the new target scene. As a result, the original assumption is incorrect, which makes it difficult for the original detector to effectively detect vehicles in the new scene. To solve this, scene adaptive learning or so-called ''transfer learning'' has gradually been introduced into the field of machine learning. In contrast with traditional statistical-learning-based methods, scene-adaptive learning focuses on learning different data distributions and using knowledge from existing scenes to help classifier learning in a new scene.

To achieve a successful transfer between the original detector and the samples in the target scene, new training samples in the target scene should firstly be obtained. High quality samples in the target scene can provide effective information for any subsequent sample transfer and classifier training process. There are two overall categories of sample generation frameworks in the existing visual vehicle detection domain. The first is manual sample labeling [8] and the second is automated sample labeling using the original classifiers [9], [10]. Due to a low degree of automation, manual sample labeling is out of use while reliability of sample selection and labeling for the latter method still needs to be improved. However, once new samples have been selected and labeled, there is still the important issue of designing the training method with scene-adaptive learning capability.

Focusing on the automated sample selection and labeling problem and the scene adaptive classifier training problem, a scene-adaptive vehicle detection algorithm based on a composite deep structure is proposed in this article. Inspired by the Bagging (Bootstrap aggregating) mechanism, multiple relatively independent source samples are firstly used to build multiple classifiers and then voting is used to generate target training samples with confidence scores. The automatic feature extraction ability of DCNN (Deep Convolutional Neural Network) is then used to perform source-target scene feature similarity calculations with a deep auto-encoder, in order to design a composite deep-structure-based scene-adaptive classifier and its training method.

In summary, the contribution of this work is with two parts. First, we add a deep auto-encoder to traditional DCNN to get a deep structure to make it have scene adaptive ability while exist deep model do not have scene adaptive ability. Second, to generate samples in target scene, we proposed a confidence samples generation with voting mechanism.

II. RELATED WORK

Due to the growing requirement for the development of intelligent transport systems and an increase of road video surveillance and on-board video, vehicle detection tasks in different scenarios is becoming increasingly important. Therefore, research into the adaptive ability of existing vehicle detectors in new scenes is receiving more academic importance and practical significance.

As mentioned before, sample selection and classifier training are the two main aspects of scene-adaptive learning. For automatic sample selection, Nair [11] has used a background subtraction method to obtain the target sample for a fixed camera. However, the robustness of the samples selected by this method is relatively low, which can easily cause a drift in the detectors. Rosenberg *et al.* [12] has proposed a self-training method using semi-supervised learning, which selects the sub-images with a high detection score as the samples in the target scene. The limitation of this method is that since only the sub images with a high score are chosen, these samples cannot fully reflect the data distribution characteristics of a new scene. Wang [13] has firstly detected objects with an original detector in a new scene and then combined text information such as motion, size, position, appearance and trajectory to calculate the confidence score of each sample. The effect of Wang's method tightly relies on prior information of the scene which has limited its applicability. Sharma *et al.* [14] has used an additional tracking process to detect objects and has chosen undetected objects as positive samples and falsely-detected objects as negative samples.

For scene-adaptive vehicle detection classifier training, Li [15] has proposed a ConvNet-framework-based training method which achieves classifier transfer by retaining the shared filters and eliminating the non-shared filters. There are not many other studies that specifically focus on vehicle-detection-based classifier transfer learning and most studies are concentrated on the field of pedestrian detection. Wang has proposed a general transfer framework: Confidence-Encoded SVM [13]. In this framework, a penalty item of the source-target scenes correlation has been added to the original SVM object function. Xian-Bin *et al.* [16] has

expanded the original Adaboost algorithm and named it the ''ITLAdaBoost''. In ITLAdaBoost, the sample weights are dynamically adjusted based on the classification error rates of the source and target data sets separately. Misclassified source sample weights are decreased and misclassified target sample weights are increased. The final classifier remains a linear combination of weak classifiers. Recently, deep model has been applied in many object detection tasks including supervised learning [17], [18] and semi-supervised learning [19], [20]. But all of the existed transfer classifier designs and training methods mentioned are based on traditional artificially-crafted features and shallow models which have not utilized the feature automatic learning ability of deep structure such as DCNN, Fast RCNN, YOLO2 and SSD.

Due to this, in our work, a deep-structure-based scene-adaptive vehicle detection method will be proposed.

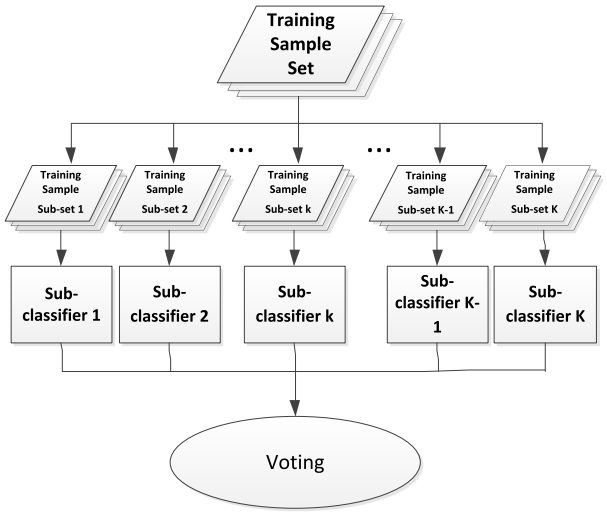


FIGURE 1. Bagging resemblance learning method.

III. CONFIDENCE SAMPLES GENERATION WITH VOTING MECHANISM

The first half of the last section has outlined some existing automatic sample selection methods. However, the samples selected by those methods inevitably contain labeling noise which gives different confidence levels of each labeled samples. Therefore, there is a requirement to find a new method to not only select the target scene samples but also be able to measure the confidence level of each sample. Bagging (Bootstrap aggregating) is a resemblance learning method which that matches different sub-classifiers to a resembling strong classifier (Fig.1). In this framework, several separate sub-training datasets are firstly prepared and each sub-classifier is trained on a single dataset only. The final output of each classifier is decided by each sub-classifier using a voting mechanism.

Inspired by the voting mechanism, we have prepared a few relatively independent source training data sets $\gamma_k (k = 1, ..., K$ and K is the number of source training data sets). All of the training data sets are captured under different weather conditions, different scenarios and even different camera equipment. Each vehicle classifier Φ_k is trained with Viola and Jones' method using a single independent source training data set γ_k. In the subsequent voting process, all vehicle classifiers Φ_k make a judgment on the unknown sample which is used to calculate a confidence score. It is assumed here that $\widehat{k}$ out of K classifiers consider the unknown sample to be the vehicle and thus the confidence scores of this sample is calculated with function (1).

$$s = c * \widehat{k}/K \tag{1}$$

In function (1), c is the upper limit of the confidence score s and its value range is between [0,1] . In our work, $K = 9$ independent source training data sets are used.

IV. SCENE-ADAPTIVE CLASSIFIER TRAINING USING COMPOSITE DEEP STRUCTURE

A. DESCRIPTION OF PROPOSED COMPOSITE DEEP STRUCTURE

Traditional scene adaptive classifiers are largely based on fixed artificial features. Thus they are only adapted at a classifier parameter level rather than a feature representation level. On the other hand, existing research results have proven that feature expression determines the upper bound of the classifier capacity and that training of the classifier can only approximate this upper bound. Therefore, it is critical to find an adaption method at a feature level. Fortunately, the recently developed deep model has the advantages of a flexible structure and the ability to feature self-learning, and thus is very suitable for our application. Therefore, a type of deep model known as the deep convolutional neural network (DCNN) is used to extract features automatically. Additionally, based on the basic structure of DCNN, a deep auto-encoder is additionally expended so that the features extracted by the DCNN can be selected with their reconstruction error in target samples. Here, the auto-encoder is used to reconstruct feature of target samples and the role of it is to measure the similarity of target-source samples in the DCNN feature space. If the reconstruct error is small for a target samples, it means that it is more similar to source samples in feature space and will be given a bigger weights in training. This composite deep structure ensures that features that are more adaptive to the target scene will be given a larger weight in training to achieve classifier transference.

A two-stage DCNN is chosen for the extraction feature vector and all the samples from the source scene and target scene will be used in this step. The reason why DCNN is applied here is due to its own structural advantages. Firstly, as one of the most common models in deep learning, DCNN is a bio-inspired architecture and learns features implicitly from the training data, enabling it to integrate feature extraction capabilities into multilayer perception by structural reconstruction and weight number reduction. Secondly, DCNN uses a special structure for local weight sharing which is

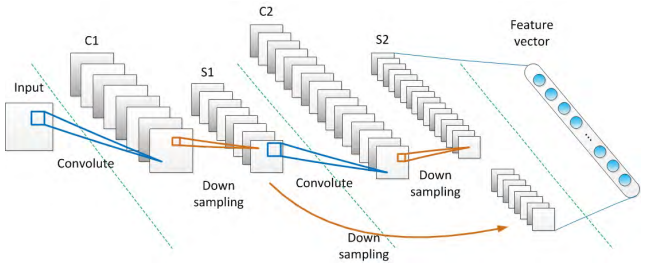


FIGURE 2. Structure and parameters of DCNN.

closer to actual biological neural networks and reduces the computational complexity of the network.

The specific structure and parameters of the DCNN used in this work is shown in Fig.2.

The designed DCNN vehicle classifier contains one input layer, two convoluted sub-sampling hidden stages and one feature vector output layer. The size of the input layer is 32×32 which is equal to the training image dimensions. The convolutional kernel of the hidden layers is 5×5 and the size of all the pools is 2×2.Max-pooling operation is chosen for sub-sampling due to its superior performance. Therefore, the size of the convolution layer and the sub-sampling layers of the two hidden stages C1, S1, C2, S2 are 28×28, 14×14, 10×10 and 5×5, respectively. The feature vector output layer contains 600 neurons and is connected to the sub-sampling layer S2 and the secondary sub-sampling layer of the sub-sampling layer S1. The feature vector is constructed here with both S1 and S2 and aims to retain the image characteristic information at multiple scales such as coarse and fine scale.

For the DCNN structure for feature generation, it is pre-trained with the samples in source dataset with the popular supervised way of random gradient descent method and the labels are put above the second DCNN layer. Beside, the initial value of the training weight parameter is selected as $[-0.05, 0.05]$ and the objective function is also cross entropy loss function.

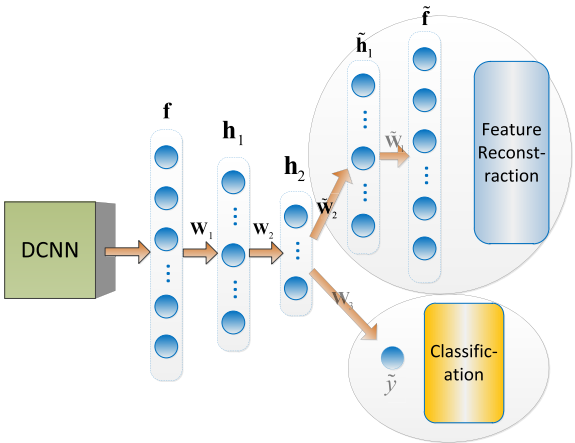


FIGURE 3. Schematic diagram of composite deep structure.

The proposed composite deep structure based on DCNN feature extraction is shown in Fig.3. As shown in this figure, f is the output feature vector of DCNN. Based on this feature vector f, two additional hidden layers h_1 and h_2, one reconstruction hidden layer $\widetilde{h}_1$, one reconstruction feature layer $\widetilde{f}$ and one classification layer y can be added. In the structure proposed above, the output feature vector f, the hidden layers h_1, h_2, $\widetilde{h}_1$ and there construction feature layer $\widetilde{f}$ essentially constitute a deep auto encoder. This deep auto encoder only receives training samples from the target scene and calculates the reconstruction error of each feature f. Here, the purpose of this auto encoder is to assess the similarity of samples between source and target scene and give different weights of the features in target training function.

The parameters are propagated by functions (2)-(6).

$$h_1 = \sigma\left(W_1^T f + b_1\right) \tag{2}$$

$$h_2 = \sigma\left(W_2^T f + b_2\right) \tag{3}$$

$$\widetilde{y} = \sigma\left(W_3^T h_2 + b_3\right) \tag{4}$$

$$\widetilde{h}_1 = \sigma\left(\widetilde{W}_2^T h_2 + \widetilde{b}_2\right) \tag{5}$$

$$\widetilde{f}_1 = \sigma\left(\widetilde{W}_1^T h_1 + \widetilde{b}_1\right) \tag{6}$$

In the functions above, $\sigma(a) = 1/\left[1\text{-}\exp(-a)\right]$ is the activation function, W_1, W_2, W_3, $\widetilde{W}_1$ and $\widetilde{W}_2$ are the weight vectors and b_1, b_2, b_3, $\widetilde{b}_1$ and $\widetilde{b}_2$ are the base vectors. These weight vectors and base vectors are all the parameters that need to be trained

B. TRAINING METHOD OF PROPOSED COMPOSITE DEEP STRUCTURE

It is assumed that the feature extracted with DCNN of the $n$tht training sample is f_n and the corresponding label of this sample is y_n. Then the parameter set of this training sample is $\{f_n, y_n, s_n, k_n\}$. Within the parameter set, $k_n = 1$ if the nth training sample is from the target scene and $k_n = 0$ otherwise. Note that s_n is the confidences core. $s_n = 1$ if the sample comes from the source scene and $s_n \in (0, 1]$ if it comes from the target scene, which can be calculated with function (1).

On the basis of this parameter set definition, the critical object function L can be designed as follow:

$$L = \sum_n e^{-\alpha L^\tau\left(f_n, \widetilde{f}_n\right)} L^c\left(y_n, \widetilde{y}_n, s_n\right) + \beta k_n L^\tau\left(f_n, \widetilde{f}_n\right) \tag{7}$$

In the object function, $L^\tau = \left(f_n, \widetilde{f}_n\right) = \left\|f_n, \widetilde{f}_n\right\|^2$ is used to calculate the reconstruction error. The feature extracted with DCNN will be reconstructed with samples from the target scene. A small reconstruction error indicates that that feature is closer to the feature distribution of the target scene. This feature will then be considered more valuable and given a larger weight. Additionally, $L^c = (y_n, \widetilde{y}_n, s_n) = s_n L^E(y_n, \widetilde{y}_n)$ where $L^E = (y_n, \widetilde{y}_n) = -y_n \log \widetilde{y}_n - (1 - y_n) \log(1 - \widetilde{y}_n)$ is a cross entropy loss function. This function is utilized to measure the difference between the estimated labels and the

Caltech 2001

MIT car data

UIUC Database

TME Motorway dataset (Sunset)

OURS Rainy

FIGURE 4. Typical images of five mentioned dataset.

actual labels. For $L^c = (y_n, \widetilde{y}_n, s_n)$, a confidence score s_n is used to further determine the difference.

Finally, the parameters of the whole composite deep structure are adjusted with the back-propagation method.

V. EXPERIMENT AND ANALYSIS

In this section, the proposed scene-adaptive vehicle detection classifier is tested on two data sets: the KITTI road

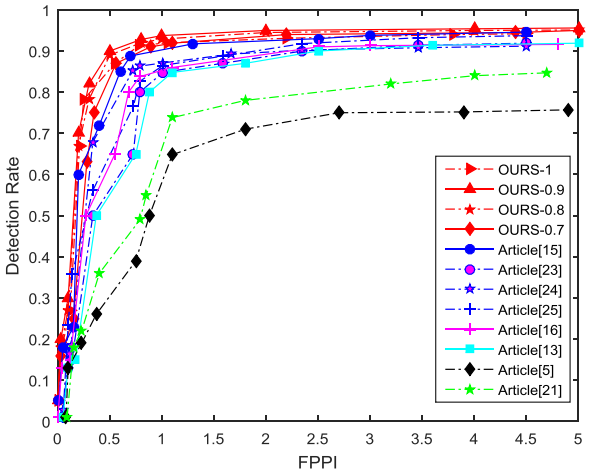


FIGURE 5. ROC curve of different methods in KITTI data set.

image data set and video sequences captured by our group. In these tests, if the coverage between the vehicle detection box and the real vehicle external rectangular box is over 80%, the vehicle detection is considered to be successfully. Based on this definition, the ROC curve is utilized as the performance evaluation method for all of the vehicle detection methods.

A. KITTI DATA SET

In this subsection, the experiments are implemented using the KITTI data set. This data set contains road images captured in a variety of road conditions and the vehicles in images are all precisely marked [22]. All images in the KITTI data set are divided into two sections, which are the training set and the test set. The training set contains 7481 images with approximately 35000 vehicles and the test set contains 7518 images with approximately 27000 vehicles.

In these experiments, the source training samples are from nine relatively independent vehicle data sets mentioned in section 3 and there are 7500 in total. In detail, the nine data sets are from six public dataset as follow: Caltech 2001, MIT car data, UIUC Database, INRIA Car Dataset, TME Motorway dataset (Daylight), TME Motorway dataset (Sunset) and three other dataset that captured by our group which are in city daytime (OURS Daytime), city rainy time (OURS Rainy) and city evening (OURS evening) respectively. some typical images of the mentioned dataset are shown in Fig.4.

The training set of KITTI is set as the target scene and used to generate the target scene positive samples. The confidence sample selection method mentioned in section 3 will be used here to generate the target scene samples with an associated confidence score. The value of c is set to 0.7, 0.8, 0.9 and 1.0 separately in the experiment. All negative training samples are constituted of images that do not contain vehicles in the KITTI set and there are 20000 in total. Finally, 2000 images containing 7218 vehicles from the KITTI test set are selected as the test set of this experiment.

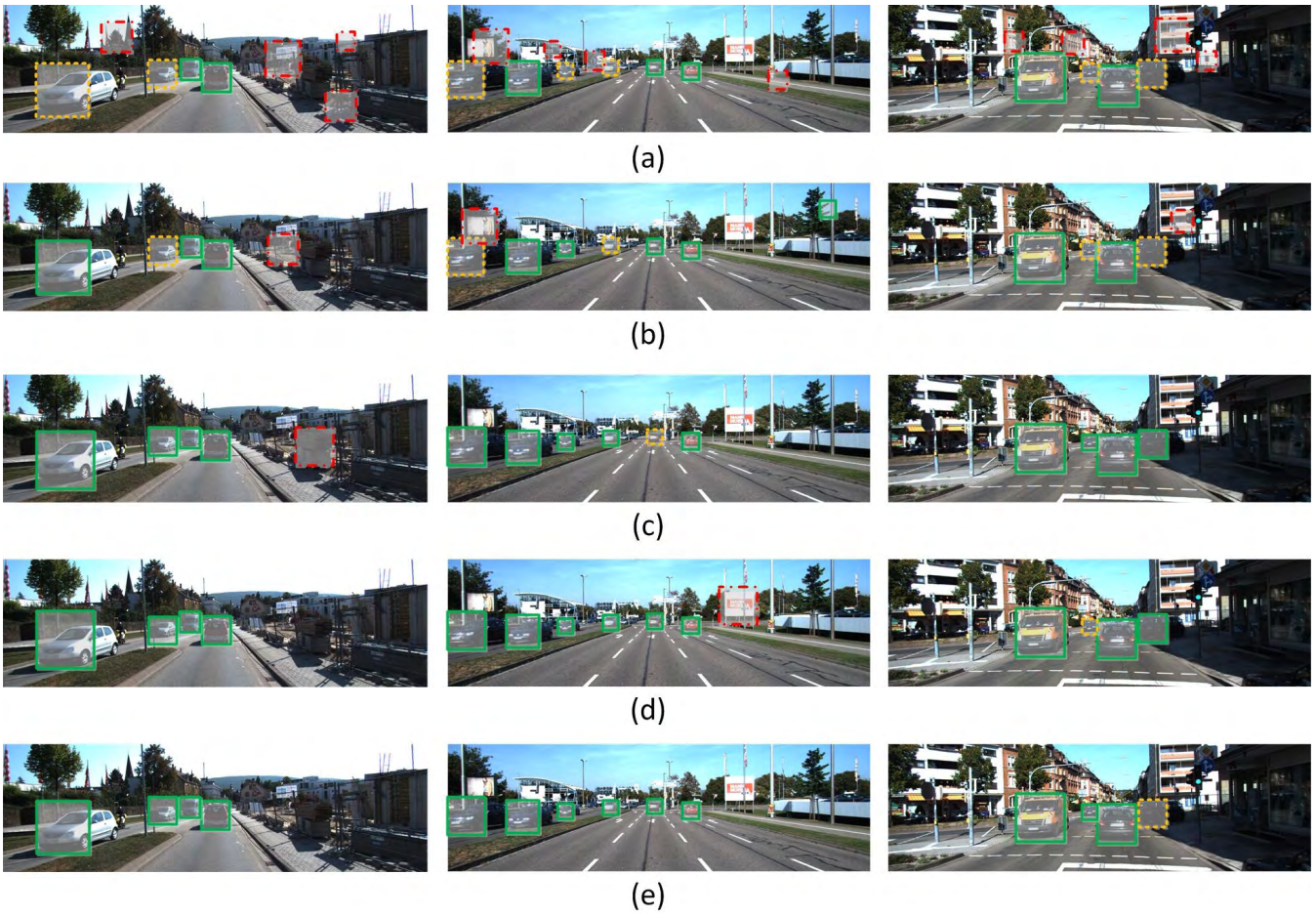


FIGURE 6. Detection results of different methods using the KITTI data set. (a) Vehicle detection results of DCNN; (b) Vehicle detection results of ConvNet; (c) Vehicle detection results of YOLO2; (d) Vehicle detection results of SSD; (e) Vehicle detection results of our method.

To assess our method, the proposed algorithm is compared with some existing state-of-the-art object recognition algorithms, including non-scene-adaptive learning algorithms and scene-adaptive learning algorithms. In this experiment, the non-scene-adaptive learning algorithms are the Cascaded Adaboost [5], DCNN [21], Faster R-CNN [23], YOLO2 [24] and SSD [25] and the scene-adaptive learning algorithms are the Confidence-Encoded SVM [13], the ITL-AdaBoost [16] and the ConvNet described in article [15]. Here, all the deep model we compared are use their original structure and they are first unsupervised trained with VOC 2017 dataset to get initial network parameters and then supervised trained by the mentioned vehicle dataset.

The ROC curve for each of these methods is shown in Fig.5. In this curve, the abscissa shows the FPPI (False Positives Per Image) of each method and the ordinate shows the corresponding detection rate. The labels OURS-1, OURS-0.9, OURS-0.8 and OURS-0.7 in the figure represent our classifier with the value of c.

It can be seen from the ROC curve that the proposed scene-adaptive vehicle detection classifier has the best detection performance of all methods when $c = 1$. When FPPI is 1, the detection rate of our method, ConvNet, ITL-AdaBoost and Confidence-Encoded SVM are 93.75%, 90.50%, 85.25% and 83.75% respectively. Additionally, when FPPI is 1, the detection rates of three types of non-scene adaptive methods are relatively low. Fig.6 (a) -Fig.6 (c) show some of the detection results of DCNN, ConvNet, YOLO2, SSD and our method using the KITTI test set. In Fig.6, a green box represents a correctly detected vehicle, a yellow dashed box represents an undetected vehicle and a red dashed box represents a falsely detected vehicle.

B. VIDEO SEQUENCES CAPTURED BY OUR GROUP

For this test, our group captured a road video on a busy road. The video is around 327 seconds long and contains 15000 frames. In the experiment, 250 frames of the first 5000 frames with an interval of 20-frames are selected for the target scene positive sample generation and the remaining 10000 frames are chosen as the test set. The experiment ROC curve is shown in Fig.7.

From the ROC curve, it can also be seen that the proposed composite deep-structure-based scene-adaptive

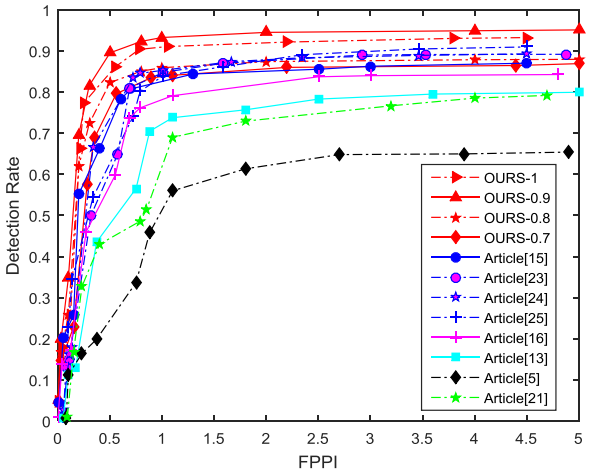


FIGURE 7. ROC curve of different methods in our video sequences data set.

vehicle detection method outperforms existing scene-adaptive object detection methods [13], [16], [15]. Furthermore, the experimental results also demonstrate that a scene-adaptive-based framework is superior to non-scene-adaptive-based methods.

Based on the experiment results, we conclude that although newly proposed deep model such as YOLO2 and SSD is superior than most of the existed work, their performance still drop significantly when the test set in target scene is significantly different from the training set in training scene, especially when training sample numbers are not very big. In the contrary, our algorithm use a small amount of automatic labeled test samples in target scene to adjust the deep model to make it better satisfy the target scene.

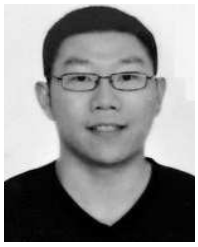
VI. CONCLUSION

This paper has proposed a scene-adaptive vehicle detection algorithm based on confidence sample generation with a voting mechanism and a composite deep structure. Inspired by the Bagging (Bootstrap aggregating)mechanism, multiple relatively independent source samples have been firstly used to build multiple classifiers and then voting is used to generate confidence scores for the target training samples. A composite deep-structure-based scene-adaptive classifier and its training method have then been designed using the automatic feature extraction ability of DCNN (Deep Convolutional Neural Network) and performing source-target scene feature similarity calculation with a deep auto-encoder. Experiments using the KITTI data set and our own data set demonstrate that this method exhibits the advantages of a high degree of automation and a high vehicle detection rate compared with existing state-of-the-art methods. The limitation of this method is that the confidence assignment method is a simple linear function dependent on the sub-classifier members which is relatively subjective and lacks a theoretical basis. Therefore, research is still needed to investigate this further.

REFERENCES

[1] L.-W. Tsai, J.-W. Hsieh, and K.-C. Fan, "Vehicle detection using normalized color and edge map," *IEEE Trans. Image Process.*, vol. 16, no. 3, pp. 850–864, Mar. 2007.
[2] D. R. Martin, C. C. Fowlkes, and J. Malik, "Learning to detect natural image boundaries using local brightness, color, and texture cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 5, pp. 530–549, May 2004.
[3] Z. Zhang, Y. Xu, J. Yang, X. Li, and D. Zhang, "A survey of sparse representation: Algorithms and applications," *IEEE Access*, vol. 3, pp. 490–530, 2015.
[4] X. Wen, L. Shao, W. Fang, and Y. Xue, "Efficient feature selection and classification for vehicle detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 3, pp. 508–517, Mar. 2015.
[5] A. Khammari, F. Nashashibi, Y. Abramson, and C. Laurgeau, "Vehicle detection combining gradient analysis and AdaBoost classification," in *Proc. IEEE Intell. Transp. Syst.*, Sep. 2005, pp. 66–71.
[6] A. Broggi, E. Cardarelli, S. Cattani, P. Medici, and M. Sabbatelli, "Vehicle detection for autonomous parking using a Soft-Cascade AdaBoost classifier," in *Proc. IEEE Intell. Veh. Symp.*, Jun. 2014, pp. 912–917.
[7] Y. Cai, Z. Liu, H. Wang, and X. Sun, "Saliency-based pedestrian detection in far infrared images," *IEEE Access*, vol. 5, pp. 5013–5019, 2017.
[8] H. Daumé, III, and D. Marcu, "Domain adaptation for statistical classifiers," *J. Artif. Intell. Res.*, vol. 26, no. 1, pp. 101–126, May 2006.
[9] M. Wang, W. Li, and X. Wang, "Transferring a generic pedestrian detector towards specific scenes," in *Proc. IEEE CVPR*, Providence, RI, USA, Jun. 2012, pp. 3274–3281.
[10] F. Liang, S. Tang, Y. Wang, Q. Han, and J. Li, "A sparse coding based transfer learning framework for pedestrian detection," in *Proc. Int. Conf. Multimedia Modeling*, Huangshan, China, 2013, pp. 272–282.
[11] V. Nair and J. Clark, "An unsupervised online learning framework for moving object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Washington, DC, USA, Jun. 2004, pp. 317–324.
[12] C. Rosenberg, M. Hebert, and H. Schneiderman, "Semi-supervised self-training of object detection models," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Breckenridge, CO, USA, Jun. 2005, pp. 29–36.
[13] M. Wang and X. Wang, "Automatic adaptation of a generic pedestrian detector to a specific traffic scene," in *Proc. IEEE CVPR*, Colorado Springs, CO, USA, Jun. 2011, pp. 3401–3408.
[14] P. Sharma, C. Huang, and R. Nevatia, "Unsupervised incremental learning for improved object detection in a video," in *Proc. IEEE CVPR*, Providence, RI, USA, Jun. 2012, pp. 3298–3305.
[15] X. Li, M. Ye, M. Fu, P. Xu, and T. Li, "Domain adaption of vehicle detector based on convolutional neural networks," *Int. J. Control, Autom. Syst.*, vol. 13, no. 4, pp. 1020–1031, 2015.
[16] X. Cao, Z. Wang, P. Yan, and X. Li, "Transfer learning for pedestrian detection," *Neurocomputing*, vol. 100, pp. 51–57, Jan. 2013.
[17] D. Zhang, J. Han, C. Li, J. Wang, and X. Li, "Detection of co-salient objects by looking deep and wide," *Int. J. Comput. Vis.*, vol. 120, no. 2, pp. 215–232, 2016.
[18] D. Zhang, J. Han, J. Han, and L. Shao, "Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1163–1176, Jun. 2016.
[19] D. Zhang, J. Han, J. Han, and L. Shao, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3660–3671, Jun. 2016.
[20] X. Lu, X. Li, and L. Mou, "Semi-supervised multitask learning for scene recognition," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1967–1976, Sep. 2015.
[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[22] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE CVPR*, Providence, RI, USA, Jun. 2012, pp. 3354–3361.
[23] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2015, pp. 1440–1448.
[24] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 779–788.
[25] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

YINGFENG CAI received the B.S., M.S., and Ph.D. degrees from the School of Instrument Science and Engineering, Southeast University, Nanjing, China, respectively. In 2013, she joined as an Assistant Professor with the Automotive Engineering Research Institute, Jiangsu University. Her research interests include computer vision, intelligent transportation systems, and intelligent automobiles.



HAI WANG received the B.S., M.S., and Ph.D. degrees from the School of Instrument Science and Engineering, Southeast University, Nanjing, China, respectively. In 2012, he joined as an Assistant Professor with the School of Automotive and Traffic Engineering, Jiangsu University. His research interests include computer vision, intelligent transportation systems, and intelligent automobiles.

ZHENGYANG ZHENG received the B.S. degree from Nantong University. He is currently pursuing the master's degree with Jiangsu University. His research interests include intelligent vehicles.

XIAOQIANG SUN received the Ph.D. degree in vehicle engineering from Jiangsu University, Zhenjiang, China, in 2016. His research interests include hybrid model predictive control and engineering application of hybrid systems theory, intelligent automobiles, and vehicle control systems.