IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 15, 2021, accepted January 27, 2021, date of publication February 8, 2021, date of current version February 16, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3057616

Pedestrian- and Vehicle-Detection Algorithm Based on Improved Aggregated Channel Features

JIE HUA[1], YING SHIA[1], CHANGJUN XIE[1], (Member, IEEE), HUI ZHANG[2], AND JIAN ZHANG[1]
[1]School of Automation, Wuhan University of Technology, Wuhan 430070, China
[2]Intelligent Transportation Systems Research Center (ITSC), Wuhan University of Technology, Wuhan 430070, China

Corresponding authors: Hui Zhang (zhanghuiits@whut.edu.cn) and Jian Zhang (jian_zhang@whut.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFB1600800, and in part by the National Natural Science Foundation of China under Grant U1764262 and Grant 51805388.

ABSTRACT In advanced driver-assistance systems (ADAS), the accuracy and real-time performance of pedestrian- and vehicle-detection algorithms based on vision sensors are crucial for safety. Here, a lightweight detection algorithm based on aggregated channel features (ACFs),consisting of a context pixel ACF (CP-ACF) pedestrian detector and a multiview ACF (Mv-ACF) vehicle detector, is proposed to rapidly and precisely understand road scenes. The former fuses local and context information to improve the robustness to pedestrian deformation, while the latter contains a number of subclass detectors to alleviate intraclass differences due to different viewing angles. Compared to the original ACF, the CP-ACF pedestrian detector reduces the average miss rate (AMR) by 6.34%. The Mv-ACF vehicle detector improves the average precision (AP) by 40.26% on average at easy, moderate and hard levels. This remarkable effectiveness is due to the spectrum clustering of multiview samples and the resulting integration of these subclass detectors via confidence score calibration, which reduces the intraclass differences of vehicles. Since feature extraction takes up 68.8% of the total detection time, a mechanism of feature sharing between pedestrian and vehicle detectors is advanced to reduce the time spent in feature extraction. A strategy based on ground-plane constraints (GPCs) is proposed to control false detection of pedestrians and vehicles by incorporating road prior information, which reduces the AMR by 1.07% for CP-ACF pedestrian detectors and improves the AP by 0.27% on average for Mv-ACF vehicle detectors. Thus, the proposed algorithm can effectively control false detection by road prior information.

INDEX TERMS Pedestrian and vehicle detection, ACF, anti-deformation, multiview, ground plane constraint, lightweight.

I. INTRODUCTION

Alongside the increase in profits in the automobile industry in recent years, the frequency of traffic accidents has grown due to various factors. Confronted with this challenge, many automobile manufacturers are developing advanced driver-assistance systems (ADAS), including various sensors and algorithms. The accuracy and real-time performance of pedestrian- and vehicle-detection algorithms based on vision sensors are crucial since safety is a top priority.

Increasing effort has been devoted to improving the accuracy and real-time performance of detection algorithms. Recently, vision-based object detection has become a research hotspot due to the compelling success of deep learning [1]–[9]. The single-shot multibox detector (SSD) and Faster region-based convolutional neural network (R-CNN) algorithms are representative of algorithms with high accuracy in detecting vehicles and pedestrians. However, since these large-scale deep learning-based models need a very large number of parameters for fine-tuning, the requirements in terms of computational resources and memory are extremely high. Therefore, it is difficult to deploy these models on resource-constrained devices, e.g., CPUs or embedded gadgets. By contrast, statistical feature-based object-detection methods are less time consuming and hence more suitable. Dollár *et al.* [10] computed different types of

The associate editor coordinating the review of this manuscript and approving it for publication was Zhongyi Guo.

feature channels in images and, together with the integrated image, extracted the features of different orders; integral channel features (ICFs) and cascade AdaBoost were used to classify ICFs for pedestrian detection. Based on the ICFs, aggregated channel features (ACFs) were proposed in [11], and their calculation is identical to that of ICFs. The difference is that a pixel lookup table is used in feature extraction to improve the average precision (AP) of pedestrian detection. Based on the ACFs, filters have been added in locally decorrelated channel features (LDCF) and other algorithms [12]–[14] during feature presentation to further improve the detection performance, but the computational complexity also increases. Furthermore, the ACF algorithm and its variations are not applicable to simultaneous pedestrian and vehicle detection, not to mention resource-constrained devices.

Compared with other statistical feature-based and deep learning-based algorithms, the ACF algorithm exhibits better real-time performance with a lower hardware requirement despite lower detection performance. For instance, a hardware detector based on ACFs designed in [15] achieved relatively high performance for pedestrian detection. Therefore, in this paper, an ACF-based algorithm is designed for target-detection applications with resource-constrained devices. Specifically, an object-detection algorithm is proposed to successively eliminate one-category constraints and to improve detection performance in road scenes. To implement lightweight pedestrian and vehicle detection, its two-class detection framework consists of a pedestrian detector and a vehicle detector, which share some common features to improve real-time performance. In addition, a strategy based on the ground plane constraint (GPC) is adopted to augment the detection performance. In particular, the two-class detection framework proposed in this paper makes the following contributions:

(1) A multiclass object-detection framework is proposed for statistical learning methods, which usually can only accomplish one-class object detection through a detection framework.

(2) A feature-sharing structure between pedestrian and vehicle detectors is proposed that can reduce the total detection time of the algorithm.

(3) The pedestrian detector based on the context information fusion method effectively handles the nonrigid deformation of pedestrians.

(4) The multiview vehicle detector reduces the intraclass differences among vehicles.

(5) By using road prior information, postprocessing via the GPC further improves the performance of pedestrian and vehicle detectors.

The rest of the paper is organized as follows. First, related work on pedestrian- and vehicle-detection methods is briefly reviewed in Section II. Then, the basic framework is introduced in Section III, and its defects are analyzed. Section IV discusses improvement strategies for the basic framework. The proposed two-class detection framework for pedestrians and vehicles is evaluated in experiments in Section V, and Section VI concludes the paper.

II. RELATED WORK

A. PEDESTRIAN DETECTION

Pedestrian detection is indispensable in ADAS and unmanned driving systems, and increasing research effort has been devoted to this area. As a representative of gradient features, a histogram of oriented gradients (HOG) feature [16] was created to capture the outline and shape of a target. Although the Haar-like wavelet feature [17] was successfully applied for face detection, its performance in pedestrian detection was not satisfactory. After fusion of the Haar-like wavelet feature with motion features in [18], a fusion of various features was deemed effective for augmenting detection performance [19]. In particular, ICFs and ACFs, which contain gradient, color and texture information, were successively proposed in [10] and [11].

In recent years, feature matching with classifiers has received more attention [20]. In [21], matching the cascade rejecter approach with HOG features greatly improved accuracy and real-time performance. Multichannel Haar-like features were used to distinguish different parts of humans in [22], and a combination of a HOG support vector machine (SVM) and Haar-Cascade was considered to improve robustness [23].

In the ACF algorithm, multiple channel features are used with AdaBoost classifiers to achieve good detection performance, but the nonrigid deformation of pedestrians is still a problem. The LDCF [14], Checkerboards [13], and non-neighboring features and neighboring features (NNNF) [12] algorithms add different filtering operations to the ACF algorithm to strengthen the feature presentation. Although the addition of these algorithms alleviates the problem of nonrigid deformation of pedestrians, the computational complexity is severely increased, and thus, it is difficult to implement these solutions on resource-constrained devices, e.g., embedded gadgets.

B. VEHICLE DETECTION

Like pedestrian detection, vehicle detection is necessary in ADAS and unmanned driving systems. Appearance information was widely used as a low-level feature in early research [24]. As a priori information, color is often used to detect visual features such as lights [25] and license plates [26]. In addition, shadows due to local light changes can result in the deformation or even the loss of a vehicle [27]. One solution is to establish a color model by establishing color components, e.g., contrast [28] and brightness [29], to identify or remove shadow areas. Edges can also be utilized to improve detection performance [30].

In recent years, local features, e.g., HOG and Haar-like features, have become prevalent in this field. HOG features were first used in pedestrian detection to capture the gradient structure with local shape features and then modified for vehicle-detection applications [16]. Kim *et al.* [31] proposed

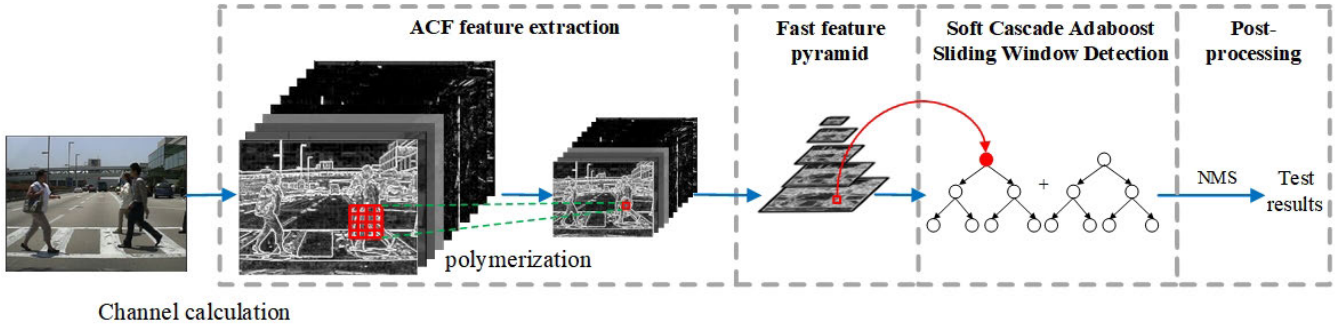


FIGURE 1. Display of some data in the line-loss dataset.

the use of HOG features to not only improve detection performance but also reduce computation time. Haar-like features were originally used in face detection [16] and were then used in [32] to capture the structure and edges in vehicle detection. Subsequently, speeded-up robust features (SURF) [33] and scale-invariant feature transform (SIFT) features [34] were created. Matching feature classifiers with local features for higher detection performance has recently become a research hotspot. For instance, Cortes and Vapnik [35] combined HOG features with SVM classifiers, and the ACF algorithm was combined with AdaBoost classifiers in [11]. In the ACF algorithm, color, brightness, gradient and other features are fused, and matching with AdaBoost classifiers greatly improves the detection performance. Although the ACF algorithm is renowned for its high accuracy in vehicle detection, its ability to discern different vehicles is still unsatisfactory. Additionally, intraclass differences due to different perspectives are a negative factor for vehicle detection.

The statistical feature-based algorithms cited above involve only a one-category classifier and are not suitable for simultaneous pedestrian and vehicle detection. In addition, it is imperative to solve other problems, such as nonrigid deformation of pedestrians, differences among vehicles and intraclass differences related to perspective. Therefore, in this paper, to implement a lightweight algorithm for simulation pedestrian and vehicle detection on resource-constrained devices, a delicate balance of the requirements for accuracy and real-time performance is considered. Specifically, the ACF algorithm is first selected as the basic framework based on its good performance in pedestrian and vehicle detection. Then, some targeted strategies are proposed to eliminate the one-category constraint to enhance its robustness to pedestrian deformation and to reinforce its ability to discern differences among vehicles and intraclass differences.

III. ACF OBJECT-DETECTION ALGORITHM

The flowchart of the ACF object-detection algorithm is shown in Fig. 1. First, from the input image, ACFs are extracted; based on these features, a multiscale fast feature pyramid is constructed, and then a soft-cascade AdaBoost classifier is used to obtain a series of bounding boxes that may contain objects by sliding-window detection on the feature pyramid. Finally, the most accurate bounding boxes for the objects are located via a postprocessing procedure, i.e., non-maximum suppression (NMS) [10]–[14], [16], [36], [37]. In this section, the ACF object-detection algorithm is introduced in detail, and its defects are analyzed.

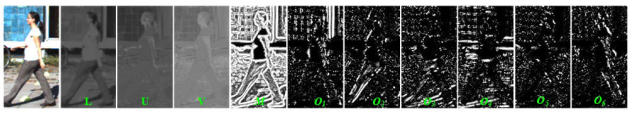


FIGURE 2. ACFs of 10 channels.

A. ACF OBJECT-DETECTION ALGORITHM

1) ACF EXTRACTION

First, the features of different channels under the same resolution are obtained via several transformations of the same input image, and $n \times n$ average pooling is performed to extract lower-resolution features. Then, a smoothing filter is applied for noise suppression. Finally, each pixel in the feature map is marked as an ACF feature. In this paper, a total of 10 channels are selected through experimental comparison [11]; the LUV color space, gradient amplitude M, and 6-direction gradient histogram Oi (i=1, 2, . . . , 6) are shown in Fig. 2.

2) CONSTRUCTION OF A FAST MULTISCALE FEATURE PYRAMID

To alleviate the sensitivity problem of single-scale ACFs brought by objects of multiple scales, a multiscale feature pyramid is established by assuming that the feature channels are correlated.

First, an image I_s is obtained by using the resampling function R and scaling image I to a scale of s; i.e.,

$$I_s = R(I, s) \tag{1}$$

Then, I_s is subjected to a linear or nonlinear transformation Ω, and the feature channel C_s at the current scale s is represented by

$$C_s = \Omega(I_s) = \Omega(R(I, s)) \tag{2}$$

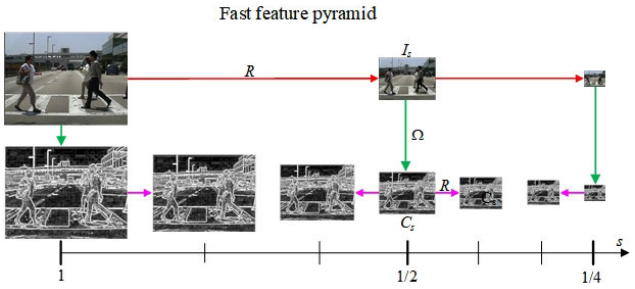


FIGURE 3. Schematic diagram of fast feature pyramid construction.

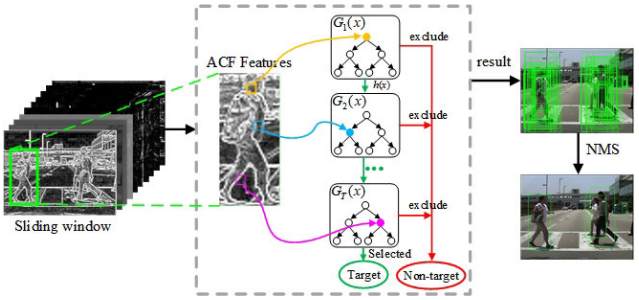


FIGURE 4. Soft-cascade AdaBoost detection.

which is derived as

$$C_s \approx R\left(C_{s'}, s/s'\right) \cdot (s/s')^{-\lambda_\Omega} \tag{3}$$

by using approximate extrapolation. In this approximate expression, λ_Ω is the approximate fixed coefficient of the channel corresponding to the transformation Ω, which can be obtained with the training data; $C_{s'}$ represents the feature channel of the scale s' closest to the scale s, as shown in Fig. 3.

3) SOFT-CASCADE AdaBoost CLASSIFIER WITH A SLIDING WINDOW

As shown in Fig. 4, to detect objects of multiple scales, a sliding window with a step size of 4 pixels is used to perform detection on all the layers of the pyramid. Within each window, the features are fed to the soft-cascade AdaBoost. Specifically, each weak classifier $G_t(x), t = 1, 2, \cdots, T$ is well trained with its output as a decision score $h_t(x)$, and then the accumulated total $H(x)$ of these decision scores is

$$H(x) = H(x) + \alpha_t \cdot h_t(x) \tag{4}$$

where α_t is the weighting coefficient of $h_t(x)$. If $H(x)$ is greater than a preset threshold, then this window is preserved for coordinate correction, from which the accurate bounding box of the object is acquired.

B. ANALYSIS OF THE ACF ALGORITHM

Although the classic ACF object-detection algorithm has higher accuracy than the algorithm based on single-channel features, the following shortcomings have hindered its application in a wider range of tasks:

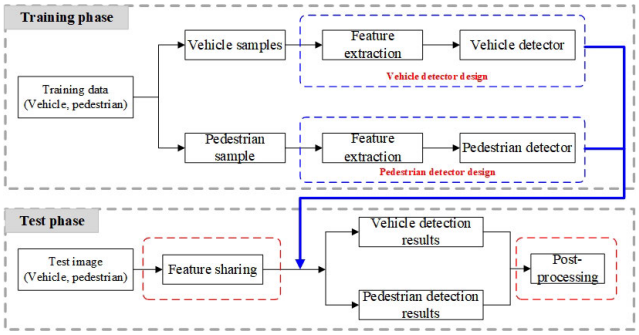


FIGURE 5. Pedestrian- and vehicle-detection framework based on feature sharing.

(1) Only one kind of object is detectable. For instance, the AdaBoost classifier is used to judge the existence of a single type of object, e.g., pedestrians or vehicles, in each candidate window. However, ADAS must detect various objects for safe driving, so the classic ACF object-detection algorithm is inadequate for this emerging technology.

(2) Its accuracy in pedestrian or vehicle detection is still unsatisfactory for practical application. For example, in the problem of pedestrian deformation, the movement and posture of a pedestrian can greatly change his or her features and render them undetectable. In addition, the feature change or intraclass difference brought by a different perspective has a significantly negative effect on vehicle detection.

(3) False detection should be further controlled. As a postprocessing tool to remove redundant detection windows, NMS uses only the features within the window, but neglecting the context information can result in false detections.

IV. PROPOSED ACF OBJECT DETECTION

Confronted with the aforementioned shortcomings, a two-class detection framework based on feature sharing is proposed.

A. TWO-CLASS DETECTION FRAMEWORK AND DATA AUGMENTATION

1) TWO-CLASS DETECTION FRAMEWORK

Feature extraction from images is generally the most time-consuming step in object detection. Traditionally, feature extraction in pedestrian detection and in vehicle detection in the same image are performed separately. Motivated by the idea that feature sharing between these two classes is expected to reduce the computation time, a pedestrian- and vehicle-detection framework is proposed. As shown in Fig. 5, this framework is divided into two phases, where ACF sharing between pedestrian and vehicle detectors can significantly improve the training efficiency. In addition, the framework structure can be readily generalized for multicategory object detection.

2) DATA AUGMENTATION

The detection accuracy can be increased by using a large amount of data. Data augmentation can produce many

FIGURE 6. Some pedestrian samples with different deformations.

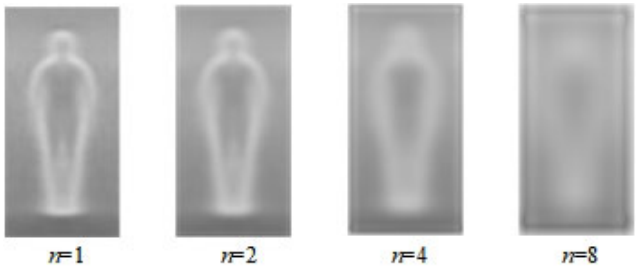


FIGURE 7. Characteristics of the gradients for different n values.

additional samples for the class in the same underlying category and is accomplished by perturbing the images of a given dataset through transformations, including flipping, rotating, cropping and scaling. In the classic ACF detection algorithm, horizontal flipping is used for data augmentation. Specifically, the training samples are first scaled for standardization. However, this flipping worsens object misalignment and can result in lower detection accuracy. In the proposed algorithm, horizontal flipping is directly replaced by multiscale data augmentation. The training images are first scaled 1.1 times horizontally and vertically, and then the object center position is normalized. This multiscale augmentation procedure is expected to improve the robustness.

B. DESIGN OF THE ANTI-DEFORMATION PEDESTRIAN DETECTOR

As illustrated in Fig. 6, pedestrian deformation brought by walking and posture is a massive challenge for the design of pedestrian detectors.

In the classic ACF algorithm, regions of size $n \times n$ are used to perform pixel pooling aggregation on feature channels, and n is a relaxation factor of robustness to pedestrian deformation. For example, the gradient amplitude channels at different values of n are compared in Fig. 7. As the value of n increases, the resolution of the extracted features decreases, and the pedestrian contour appears increasingly vague. With respect to local features, since the value of n is fixed a priori, missed detections will occur if the robustness of the detector to large pedestrian deformations is inadequate. On the one hand, as the value of n gradually increases, low-resolution features can result in false detections. On the other hand, more contextual information is included, and the robustness to deformation increases.

Since n is constant in the classic ACF algorithm, the context information may be insufficient in some cases, resulting in weak deformation robustness. In this paper, context pixel ACFs (CP-ACFs) are proposed. As shown in Fig. 8, 2×2 average pooling is first performed on the original 10 channels

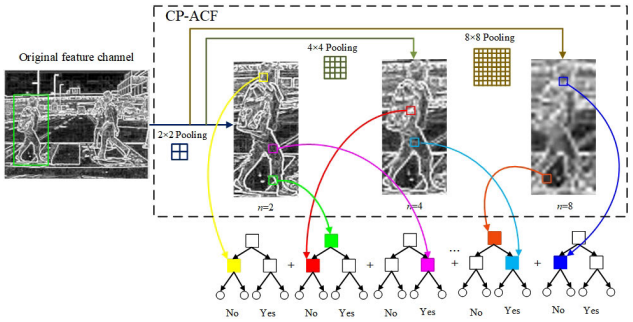


FIGURE 8. CP-ACF extraction and classification process.

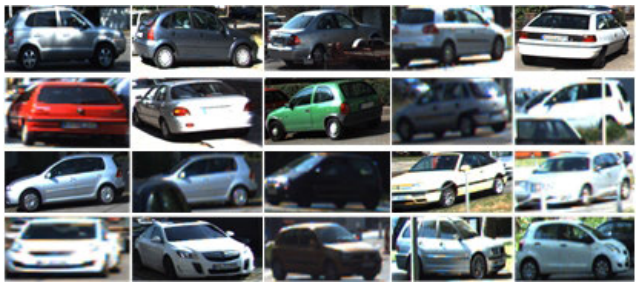


FIGURE 9. Vehicle images from different perspectives.

to obtain the ACF F_2. Then, the ACF features $F_{4\times4}$ for the 4×4 regions are obtained via 2×2 average pooling on $F_{2\times2}$, and the same procedure is repeated for the case of $F_{8\times8}$. Then, all of the $F_{4\times4}$ and $F_{8\times8}$ are upsampled to match the resolution of $F_{2\times2}$ so that these ACFs can be aggregated into a CP-ACF for all 30 channels to fuse the local and context features. Finally, the soft-cascade AdaBoost classifiers can adaptively choose the fused features of different regions in the CP-ACF channels and strengthen the robustness to pedestrian deformations.

C. DESIGN OF A MULTIVIEW VEHICLE DETECTOR

As shown in Fig. 9, large feature differences result from various vehicles or different perspectives of the same vehicle. In the classic ACF algorithm, a vehicle detector is trained with images from all perspectives, but the difference among them cannot be thoroughly captured; therefore, its detection accuracy is not satisfactory. To solve this problem, a multiview ACF (Mv-ACF) vehicle detector is proposed. Each perspective provides a subclass detector consisting of a feature extractor and a classifier. The Mv-ACF detector will undergo training and testing phases.

1) TITLE

The training process of the Mv-ACF vehicle detector is shown in Fig. 10. The training samples are first clustered according to the perspective, and then each subclass classifier is trained with the extracted features from the corresponding perspective. Specifically, the first step is realized by clustering the ACFs of these samples. Considering the very large number of ACFs, an unsupervised learning algorithm, K-means

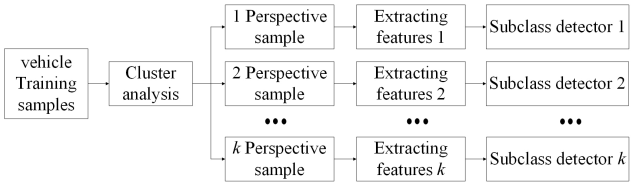


FIGURE 10. The training process of the Mv-ACF vehicle detector.

clustering, is used for classification. Due to its possible cluster degradation, we use spectral clustering (SC) [12] as follows. 1) The similarity correlation matrix among the samples is calculated, and then the feature vectors in a new feature space are obtained via spectral decomposition of the matrix. 2) K-means clustering is performed on those vectors.

2) TESTING BASED ON CONFIDENCE SCORE CALIBRATION

Similar to the classic ACF algorithm, the ACFs and the feature pyramid are shared by all the subclass detectors in the Mv-ACF detector during the testing process. Since the subclass detectors are trained with images from different perspectives, confidence scores with different distributions and bounding boxes with inconsistent geometric features, e.g., aspect ratios, are obtained. Directly merging these results will add noise and result in NMS instability and accuracy loss.

To solve this problem, the confidence scores are calibrated to rationalize their distributions [24], [25]. Assume that $Deti = \{d_{i1}, d_{i2}, \ldots, d_{ij}, \ldots, d_ir\}$ denotes the r detection results of the $i-th$ subclass detector and $dij = \{R_{ij}, c_{ij}\}$ represents the $j-th$ detection result consisting of a bounding box R_{ij} and a confidence score c_{ij}. The set $mDet_i = \{md_{i1}, md_{i2}, \ldots, md_{iu}, \ldots, md_{ir}\}$ is the calibration result, where $md_{iu} = \{R_{iu}, c'_{ij}\}$. The purpose of confidence score calibration is to make $c'_{ij} = g_i(c_{ij}) \in [0, 1]$ via a calibration function g_i. We need to ensure that g_i is an increasing function and that the calibrated confidence score can be used for classification problems. In this paper, the following parametric logistic regression is used to normalize the score [26]:

$$c'_{ij} = \frac{1}{1+\exp\ (A_i \cdot\ c_{ij} + B_i)} \tag{5}$$

where the parameters Ai and Bi of the $i-th$ subclass detector are obtained by minimizing the regularization maximum likelihood

$$\arg\min_{A_i,B_i} \ -\sum_{j=1}^{r} [t_j \log c'_{ij} + (1-t_j)\log(1-c'_{ij})] \tag{6}$$

By substituting (5) into (6), we obtain

$$\arg\min_{A_i,B_i} \sum_{j=1}^{r} [(t_j-1)(A_i \cdot c_{ij}+B_i)+\log(1+\exp(A_i \cdot c_{ij}+B_i))] \tag{7}$$

where $t_j = \begin{cases} \frac{r_++1}{r_++2}, & y_j = +1 \\ \frac{1}{r_-+2}, & y_j = -1 \end{cases}$. Here, r_+ and r_- are the numbers of positive and negative images, respectively, in the $i-th$ subclass for training parameters Ai and Bi, yj is the label of the $j-th$ sample, $y_j = +1$ represents the object, and $y_j = -1$ represents the background.

FIGURE 11. Illustration of false pedestrian detection.

Regarding the feature-sharing strategy for the pedestrian- and vehicle-detection framework in Fig. 5, the same 10 original ACF channels are used for both the Mv-ACF and CP-ACF detectors. For the Mv-ACF detector, a 2×2 average pooling operation is used for feature extraction, while average pooling sizes of 2×2, 4×4, and 8×8 are successively utilized to extract features for the CP-ACF detector. Therefore, the features used by Mv-ACF are included in those extracted by the CP-ACF detector. Since the construction procedure of the feature pyramid is unchanged, the feature pyramid used by the CP-ACF detector can be shared with the Mv-ACF detector. In addition, k Mv-ACF subclass detectors can also directly share that pyramid, making the final detection framework more efficient.

D. FALSE-DETECTION CONTROL STRATEGY BASED ON THE GPC

During sliding-window detection, as a postprocessing tool to remove redundant detection windows, NMS adopted by the classic ACF algorithm considers only features within that window and ignores other context information, which is a major source of false detection. As shown in Fig. 11, this strategy considers only the visualization of the yellow bounding boxes but cannot effectively eliminate the false-detection objects marked by red circles.

The road prior information is helpful for controlling false detection of pedestrians and vehicles. For the road scene, the height H of the bounding box for 12,186 pedestrians and 15,891 vehicles in the Caltech [19] and KITTI [38] datasets and the lower edge position coordinate Y of that bounding box are calculated. Fig. 12 shows the relationship between H and Y in the road scene, which conforms to the following statistical law:

$$H = f(Y) \tag{8}$$

Accordingly, a simple control strategy for false detection via the GPC is proposed. Concretely, for a candidate object, if the relationship between H and Y cannot be fit by the above equation, then a false detection of this object is judged.

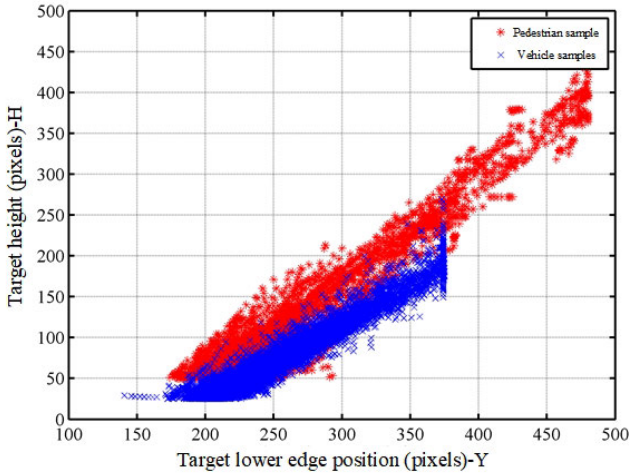


FIGURE 12. Statistics of the object height and lower edge position.

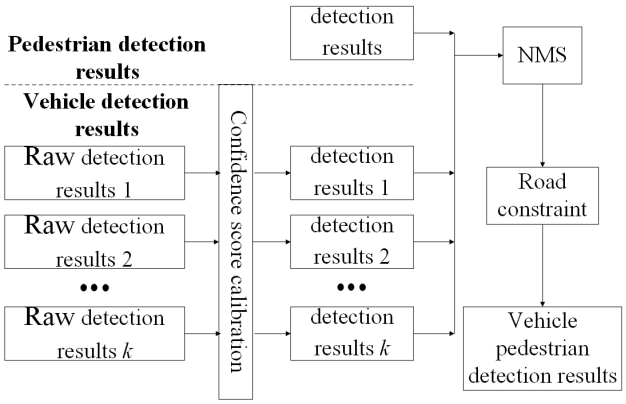


FIGURE 13. Postprocessing in the pedestrian- and vehicle-detection algorithm.

A regression model W is used to describe the statistical relationship f between H and Y and can be obtained by using an SVM to train the normalized H and Y of the samples. During testing, we assume that a bounding box after NMS is $\{x, y, w, h\}$, its lower edge position is $y + h$, the trained regression model W is used to calculate h', and the relative error between h and h' is defined as

$$E = abs(\frac{h - h'}{h'}) \tag{9}$$

where abs(Â·) represents the absolute value operator. If the value of E exceeds a preset threshold, then this bounding box is marked as a false detection; otherwise, it is a correct detection. The whole postprocessing procedure is shown in Fig. 13. We first use K subclass detectors to obtain the vehicle-detection results at K angles and calibrate the confidence scores. Then, both the vehicle-detection and pedestrian-detection results are processed via NMS. Finally, we use the road-constraint strategy to modify the detection results and obtain the final vehicle- and pedestrian-detection results.

V. EXPERIMENT AND DISCUSSION

In this section, the dataset and the evaluation criteria are first introduced, and then a comparative study is performed to verify the effectiveness of the proposed strategies. All experiments are supported by an Intel(R) Xeon(R) E5-2620 v4 CPU and 32 GB of memory.

A. DATASETS AND EVALUATION CRITERIA

Datasets play a critical role in object detection since they provide a benchmark for comparing competing algorithms. Well-known datasets for pedestrian detection are the Caltech and KITTI datasets, which contain pictures from noisy environments, e.g., foggy, rainy and snowy conditions.

In the Caltech dataset, the training set and the testing set contain 6 videos and 5 videos, respectively, at a resolution of 640×480. To expand the training set, the sampling frequency of the videos is increased threefold, and 12,823 images are acquired. Similarly, the testing set includes 4024 images. Furthermore, the miss rate (MR), false positives per image (FPPI) and average MR (AMR) are used to evaluate the performance of the detectors.

KITTI is a well-known dataset for traffic scene analysis. The training set and testing set contain 7481 and 7518 images, respectively, at a resolution of 1242×375, while the labels of the testing set are not officially released. Therefore, in this paper, the training set is randomly divided into 10 parts, 9 of which are used for training and the other for testing. In addition, the object labels in the KITTI dataset are grouped into easy, moderate and hard levels based on the extent of occlusion and truncation. In this paper, moderate-level images are used for training, and then the images at all levels are tested. The precision (P), recall (R) and AP are used as the evaluation metrics.

The pedestrian detectors considered in this paper are evaluated based on the Caltech and KITTI datasets, but only the KITTI dataset is used to assess the performance of the vehicle detector as well as the proposed two-class detection framework with the above evaluation criteria since the Caltech dataset has no vehicle labeling.

B. EXPERIMENT ON THE PEDESTRIAN DETECTOR

1) TRAINING PARAMETER CONFIGURATION

An analysis of the Caltech dataset shows that the aspect ratio of the bounding boxes of pedestrians is approximately 0.41. Accordingly, the aspect ratio of the detection model is fixed to 0.41, and the model size is fixed to 41×100 for the pedestrian detector. To obtain more context information in vehicle detection, the sliding-window size is set to 64×128 by padding the model size, and its step size is 4. The false detections at the current epoch are selected as the negative training samples in the next epoch. During the training of 4 epochs, hard-example mining is performed for 3 epochs. The number of weak classifiers at each epoch is alternately 64, 256, 1024 and 4096. The same configuration is used for the KITTI dataset except that the model size and sliding-window size are set to 20.5×50 and 32×64, respectively.

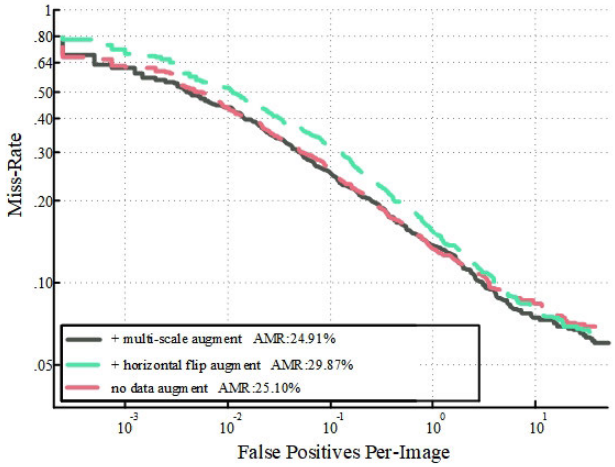


FIGURE 14. The experimental results of data augmentation.

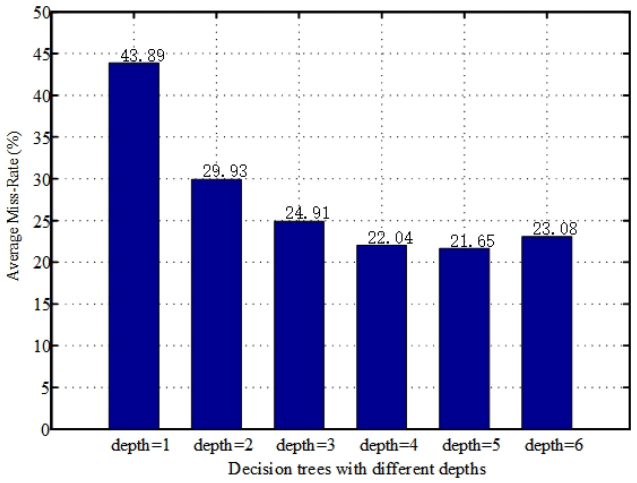


FIGURE 15. AMRs of decision trees with different maximum depths.

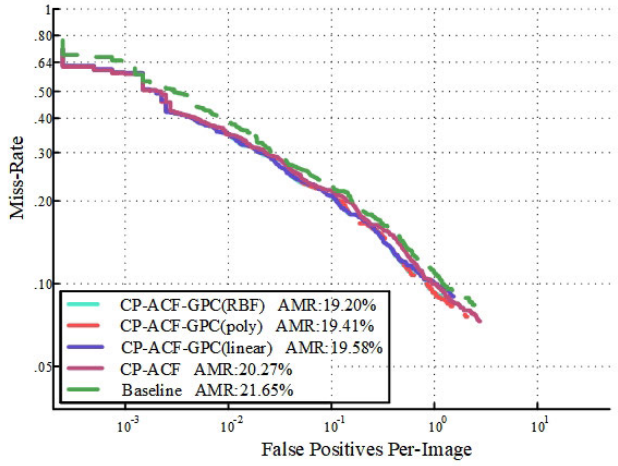


FIGURE 16. CP-ACF and GPC experimental results.

2) VERIFICATION OF THE DATA-AUGMENTATION STRATEGY

To investigate the impact of the proposed data-augmentation strategy, the pooling block size of the ACFs and the depth of the decision tree are set to 4 × 4 and 3, respectively. The relationships between the MR and FPPI in three cases, i.e., no augmentation, horizontal flipping augmentation and the proposed multiscale augmentation, are shown in Fig. 14. First, the MR decreases as the FPPI increases. Specifically, when the criterion is strict for selecting the bounding box of an object from candidate boxes, fewer false positives are accompanied by a high MR; i.e., these two evaluation metrics are negatively correlated. Second, a general increase in the MR of the classic ACF algorithm is brought by horizontal flipping augmentation, and the AMR increases from 25.1% to 29.87%, which conforms to the effect described in Section IV. By contrast, it can be inferred that the robustness is reinforced when the proposed multiscale augmentation is adopted since the MR is the lowest and the AMR is 24.91%. Therefore, this strategy will be used as the default in the following experiments.

3) DETERMINING THE BASELINE

To study the effectiveness of the proposed algorithm, the structure and parameters of the baseline model need to be determined first. The AMRs of six decision trees with depths of 1, 2, 3, 4, 5, and 6 are compared in Fig. 15. The AMR first gradually decreases as the depth increases, but the trend is reversed when the depth exceeds 5 due to saturation; i.e., the lowest AMR is obtained when the depth is 5. Therefore, the ACF model containing a decision tree with a depth of 5 is chosen as the baseline model.

4) THE VALIDITY OF THE PROPOSED STRATEGIES

(1) The validity of the CP-ACF detector. To improve the robustness to nonrigid body deformation, based on the baseline model, pixel aggregations of 2 × 2 and 8 × 8 average pooling are taken to form the CP-ACF detector. The CP-ACF detector has 30 channels and uses the same parameters as the baseline model, and its performance is shown in Fig. 16. The AMR of the CP-ACF detector is 20.27%, 1.38% lower than that of the baseline model. Therefore, the fusion of different ranges of context and local information is helpful for improving the robustness to nonrigid body deformation.

(2) The validity of the GPC. To verify the effectiveness of the GPC in controlling false detection, an SVM model is trained by using three kernel functions, namely, linear, polynomial and radial basis function (RBF) kernels, and then the postprocessing of the GPC is applied. If the relative error exceeds the preset threshold, whose optimal value is 0.38 according to an empirical test, then the corresponding bounding box is false; conversely, if the relative error is less than the preset threshold, then the corresponding bounding box is true. The performance comparison is shown in Fig. 16. All three kernel functions improve the accuracy of CP-ACF, reducing AMR by 0.69%, 0.86%, and 1.07%. The RBF is used by default in the following experiments because it outperforms the others.

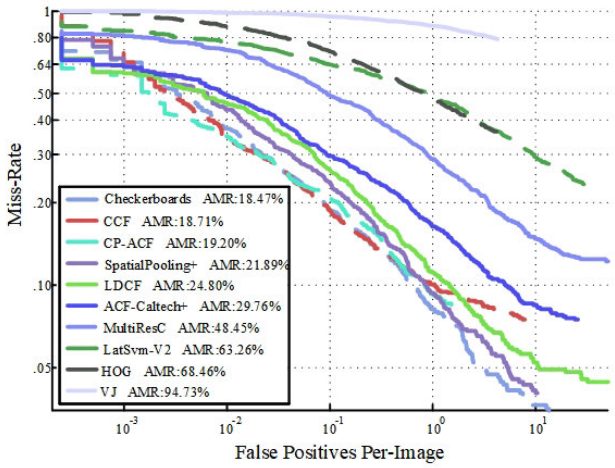


FIGURE 17. Caltech test set algorithm comparison results.

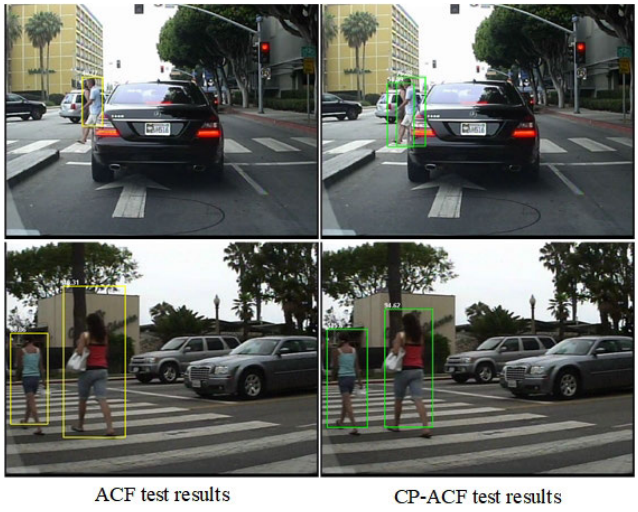


FIGURE 18. Evaluation of the pedestrian detectors on the Caltech dataset.

TABLE 1. AP at different levels of the KITTI validation set.

Pedestrian Detector	Easy	Moderate	Hard
ACF	62.1%	56.6%	48.9%
CP-ACF	**73.5%**	**64.7%**	**55.5%**

5) COMPARATIVE STUDY OF PEDESTRIAN DETECTORS

The proposed CP-ACF pedestrian detectors are compared with other mainstream pedestrian detectors based on statistical features, such as Viola-Jones (V-J) [36], HOG [16], LatSvm-V2 [37], ACF [11], LDCF [14], SpatialPooling+ [39], convolutional channel features (CCF) [40], and Checkerboards [25]. These detectors are evaluated on the Caltech dataset, and the comparison is shown in Fig. 17. V-J, HOG and LatSvm-V2, whose AMRs are too high, are not suitable for unmanned driving or ADAS. Compared with the ACF and LDCF detectors, the CP-ACF detector reduces the AMR by 10.56% and 5.6%, respectively. Although the AMR of the CP-ACF detector is 0.73% higher than that of the high-precision Checkerboards, it is less time consuming.

Fig. 18 shows the performance comparison between the proposed ACF detector and the classic ACF detector on the Caltech dataset. The former provides more accurate bounding boxes for pedestrians and even detects a pedestrian missed by the latter. Table 1 presents the comparison of ACF and CP-ACF detectors on the KITTI dataset. Compared with ACF, CP-ACF with the GPC improves the AP by 11.4%, 8.1%, and 6.6% for the easy, moderate and hard levels, respectively. It can be concluded that the CP-ACF detector outperforms the ACF detector on both datasets.

C. VEHICLE DETECTOR EXPERIMENT

1) TRAINING PARAMETER CONFIGURATION

The Mv-ACF detector clusters vehicle samples into k subclasses according to the number of perspectives and allocates each subclass an individual training subset. Then, for each training subset, the aspect ratio b of the detection model is set to the median of the aspect ratios of all samples. By fixing the height to $h = 48$, the size (w, h) of the vehicle-detection model can be obtained by $w = h\tilde{A} - - - b$, and the sliding window is padded as $(w + w/8, h + h/8)$. When k is large, the training for the detection models with default configurations in these training subsets becomes extraordinarily time consuming. Therefore, the maximum depth of the involved decision tree is set to 2. During the training of 4 epochs, bootstrapping is used for 3 epochs. The number of weak classifiers at each epoch is alternately 32, 128, 512, and 2048. All detectors for each perspective detector share the same parameters except the training data and model size.

2) DETERMINING THE BASELINE

To verify the effectiveness of the proposed algorithm, the structure and parameters of the baseline model need to be determined. SC is used to divide the training samples into 6 groups with $k = 1, 5, 10, 15, 20, 25$, where $k = 1$ represents the original ACF detector, which is chosen as the baseline.

3) THE VALIDITY OF THE PROPOSED STRATEGIES

(1) Effect of the number of subclasses k. To find an optimal number of subclasses for the Mv-ACF detector to achieve satisfactory performance, SC is adopted to divide the training set into k subsets, which are then used to train the corresponding detectors. The PR and AP values when $k = 1, 5, 10, 15, 20, 25$ are compared in Fig. 19.

It can be seen that the AP roughly increases as k increases at the easy, moderate and hard levels, while this trend changes when $k = 25$. The reason for this phenomenon is that when the number of subclasses is large, the number of samples in each training subset becomes insufficient, which results in a decline in accuracy. In addition, the detection slows when k is large. Among these 6 values of k, the optimal values of the AP at the 3 levels, which are 82.9%, 78.4%, and 60.7%, respectively, are all obtained when $k = 20$. It is inferred that the intraclass difference in a vehicle can be sufficiently represented from 20 perspectives.

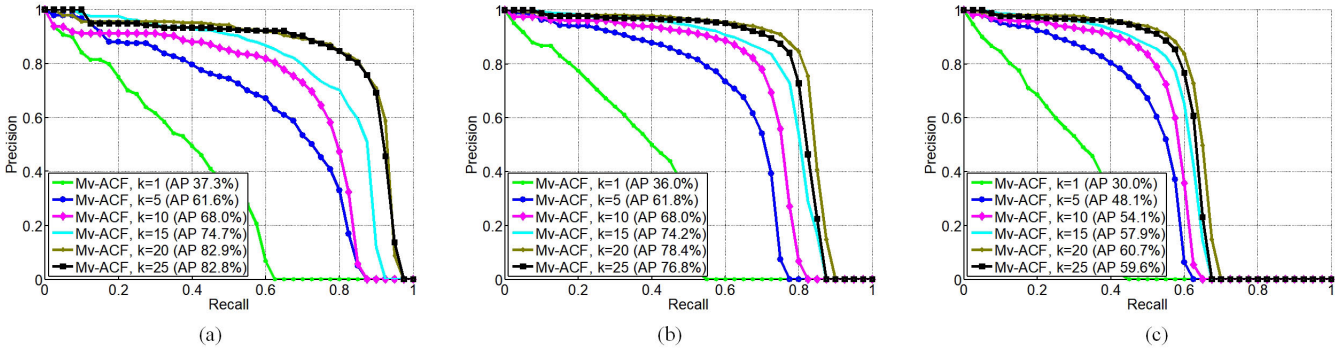


FIGURE 19. Comparison of the PR and AP values for different numbers of subclasses.

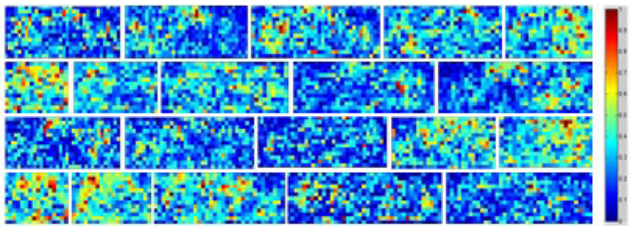


FIGURE 20. Visualization of 20 subclass detectors.

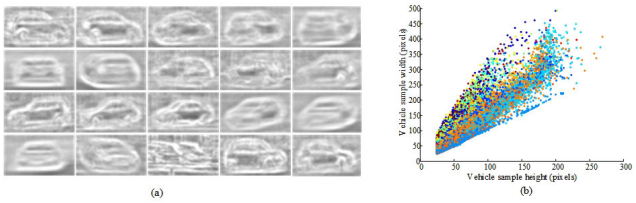


FIGURE 21. (a) Average gradient channels of different subclasses; (b) relation between the width and height of different subclass samples.

Therefore, we choose $k = 20$ in the following experiment. Fig. 20 shows the position weights from the perspective of 20 subclass detectors. The larger the weight of a position, the more attention is paid to its feature by the detector from the corresponding perspective.

(2) The validity of SC. By using SC, the samples are clustered into 20 categories to train the Mv-ACF detector, which is compared with its counterpart that adopts classic K-means clustering. The comparison of the AP values at the easy, moderate and hard levels is shown in Table 2. The AP values obtained by SC at the three levels are 0.9%, 1.9% and 1.4% higher, respectively, than those obtained by classic K-means clustering because SC can map the ACFs of samples to more classifiable features via spectral decomposition of the correlation matrix. In addition, the 15,891 vehicle samples in the KITTI dataset are used to visualize the SC. As shown in Fig. 21a, the average gradient channel for each subclass exhibits different appearance characteristics. Fig. 21b shows scatter plots of the samples with different heights and widths, where each subclass is marked with an individual color. The aspect ratios of the samples in the same subclass are similar. Therefore, according to the perspective, all of the samples can be clustered into different subclasses to effectively mitigate the intraclass differences among vehicles.

(3) The validity of the GPC and confidence score calibration. The effectiveness of the proposed GPC and confidence score calibration with the SC strategy is verified in the case of $k = 20$. The comparison results are shown in Table 2. When the GPC strategy with a threshold of 0.35 is used, the AP increases by 0.5%, 0.2% and 0.1% at the easy, moderate and hard levels, respectively. Then, via postprocessing of the confidence score calibration, the AP further increases by 1.2% at the easy level, remains the same at the moderate level and increases slightly at the hard level.

TABLE 2. AP at different levels of KITTI.

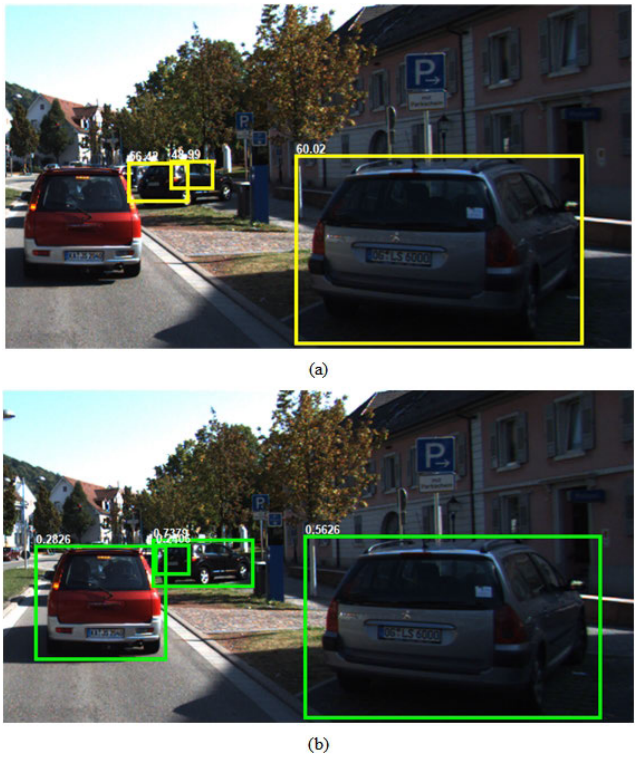
Algorithm	Easy	Moderate	Hard
ACF (k=1)	37.3%	36.0%	30.0%
Mv-ACF (K-means, k=20)	82.0%	76.5%	59.3%
Mv-ACF (SC, k=20)	82.9%	78.4%	60.7%
Mv-ACF (SC, k=20) +GPC	83.4%	78.6%	60.8%
Mv-ACF (SC, k=20) +GPC + Confidence Score Calibration	**84.6%**	**78.6%**	**60.9%**

The aforementioned strategies, i.e., SC, multiview ACF detection, GPC and confidence score calibration, all contribute to improved accuracy for vehicle detection, and SC is the top-ranked contributor.

4) COMPARATIVE STUDY OF VEHICLE DETECTORS

Table 2 also shows that the proposed Mv-ACF vehicle detector significantly improves the AP by 47.3%, 42.3%, and 30.7% at the three levels, respectively, compared with the classic ACF detector because the clustering strategy of multiperspective samples effectively solves the problem of intraclass differences.

Fig. 22 illustrates some detection results for the KITTI dataset. The scores of the bounding boxes are normalized after confidence score calibration, which facilitates the subsequent analysis. Superior to the ACF detector, the Mv-ACF detector successfully discovers more vehicles with different perspectives.



(a)

(b)

FIGURE 22. Test results of some vehicles in the KITTI dataset: (a) ACF test results; (b) Mv-ACF test results.

TABLE 3. Stage time analysis of the KITTI test set.

Algorithm	Feature extraction	Vehicle detection	Pedestrian detection	Total time
Proposed	0.285 s	0.066 s	0.063 s	**0.414 s**

D. TWO-CLASS DETECTION EXPERIMENT

1) THE VALIDITY OF FEATURE SHARING

To meet the real-time requirements of pedestrian and vehicle detection, a strategy of feature sharing between pedestrian and vehicle detectors is proposed. In the experiment based on the KITTI dataset, the evaluation metric is the detection time. The time spent in every step is shown in Table 3. The total time of the proposed two-class detection approach is 0.414 s, 68.8% of which is consumed by feature extraction. Therefore, the feature-sharing strategy can greatly reduce the detection time, and this characteristic will be further highlighted in the case of multiple-object categories. In addition, compared with the current mainstream neural networks, the algorithm presented in this paper has obvious speed advantages, whether compared with the one-stage object-detection network or the two-stage object-detection network, as shown in Table. 4. In particular, the proposed algorithm is 44.1% faster than You Only Look Once version 3 (YOLOv3), a one-stage object-detection network known for its speed. With the same hardware configuration, the speed advantage of the proposed detection algorithm is remarkable, and it is particularly significant on resource-constrained devices.

TABLE 4. Real-time statistics of different algorithms in the KITTI test set.

Algorithm	Total time
The proposed	**0.414 s**
YOLOv3	0.74 s
SSD-300	2.20 s
FCOS_r50	7.57 s
RetinaNet_r50	8.59 s
PISA_retinanet_r50	8.66 s
RetinaNet_r101	10.67 s
PISA_RetinaNet_r101	11.95 s
Faster R-CNN	16.0 s

FIGURE 23. Two-class detection results on KITTI.

2) VISUALIZATION OF THE DETECTION RESULTS

Some detection results of the proposed algorithm on the KITTI dataset are illustrated in Fig. 23. All of the detected pedestrians and vehicles are marked with yellow and green frames, respectively, and every missed object is manually marked with a red ellipse. The proposed algorithm can simultaneously detect vehicles and pedestrians. Since all neighboring objects can be detected, the proposed algorithm is applicable in unmanned driving systems or ADAS.

VI. CONCLUSION

In this paper, simultaneous pedestrian and vehicle detection based on the ACF algorithm is studied for its application in resource-constrained devices. To eliminate the one-category constraint of the ACF algorithm, a multicategory object-detection framework is proposed that consists of a CP-ACF pedestrian detector and an Mv-ACF vehicle detector. The former fuses local and context information to improve the robustness to the deformation of pedestrians, and the latter contains a number of subclass detectors to alleviate intraclass differences due to different perspectives. SC is used to determine the number of subclasses, and the results of these subclass detectors are integrated via confidence score calibration. A mechanism of feature sharing between the pedestrian and vehicle detectors is advanced to reduce the time spent in feature extraction. A strategy based on the GPC is proposed to control false detection of pedestrians and vehicles by incorporating road prior information.

By using the proposed multiscale augmentation, the AMRs of the classic ACF and CP-ACF detectors are reduced by 4.96% and 1.38%, respectively. In terms of the AMR, the CP-ACF pedestrian detector outperforms the V-J, HOG, LatSvm-V2, ACF, LDCF, SpatialPooling+ and CCF algorithms. By adopting the proposed SC and confidence score calibration methods, the Mv-ACF vehicle detector improves

the AP by 39.57% and the AMR by 0.43% on average at the easy, moderate and hard levels compared with the baseline. This improvement is mainly achieved by the clustering strategy of multiperspective samples, which effectively deals with intraclass differences among vehicles. After postprocessing via the GPC, the AMRs of the CP-ACF pedestrian detector and Mv-ACF vehicle detector are further reduced by 1.07% and 0.27%, respectively, on average at the three levels. The evaluation of feature sharing between pedestrian and vehicle detectors shows that feature extraction takes up 68.8% of the total detection time, and thus, the proposed mechanism saves a significant amount of time. The proposed detection algorithm is confirmed to be more applicable to resource-constrained devices than the current mainstream neural networks due to its speed advantage. Promising results that meet the requirements of ADAS or unmanned driving are achieved in terms of accuracy and real-time performance with hardware design or GPU hardware acceleration, whereas the detection of small objects more accurately is a challenge that remains for future work.

REFERENCES

[1] 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. (CVPR)*, Jun. 2016, pp. 779–788.
[2] W. Liu, D. Anguelov, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
[3] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6517–6525.
[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[5] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2018, pp. 502–511.
[6] X. Hu, X. Xu, Y. Xiao, H. Chen, S. He, J. Qin, and P.-A. Heng, "SINet: A scale-insensitive convolutional neural network for fast vehicle detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 3, pp. 1010–1019, Mar. 2019.
[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, Feb. 2020.
[8] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 9627–9636.
[9] Y. Cao, K. Chen, C. C. Loy, and D. Lin, "Prime sample attention in object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11583–11591.
[10] P. Dollar, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 1–11.
[11] P. Dollar, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 8, pp. 1532–1545, Aug. 2014.
[12] J. Cao, Y. Pang, and X. Li, "Pedestrian detection inspired by appearance constancy and shape symmetry," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1316–1324.
[13] S. Zhang, R. Benenson, and B. Schiele, "Filtered channel features for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1751–1760.
[14] W. Nam, P. Dollar, and J. Han, "Local decorrelation for improved pedestrian detection," in *Proc. Neural Inf. Process. Syst.*, 2014, pp. 424–432.
[15] H. Song, B. Jeong, H. Choi, T. Cho, and H. Chung, "Hardware implementation of aggregated channel features for ADAS," in *Proc. Int. SoC Design Conf. (ISOCC)*, Oct. 2016, pp. 167–168.
[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2005, pp. 886–893.
[17] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. CVPR*, Dec. 2001, p. 1.
[18] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, 2003, pp. 734–741.
[19] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.
[20] S. Munder and D. M. Gavrila, "An experimental study on pedestrian classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1863–1868, Nov. 2006.
[21] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 1491–1498.
[22] S. Zhang, C. Bauckhage, and A. B. Cremers, "Efficient pedestrian detection via rectangular features based on a statistical shape model," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 763–775, Apr. 2015.
[23] P. Govardhan and U. C. Pati, "NIR image based pedestrian detection in night vision with cascade classification and validation," in *Proc. IEEE Int. Conf. Adv. Commun., Control Comput. Technol.*, May 2014, pp. 1435–1438.
[24] O. Barnich and M. Van Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1709–1724, Jun. 2011.
[25] D.-Y. Chen, Y.-H. Lin, and Y.-J. Peng, "Nighttime brake-light detection by Nakagami imaging," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 4, pp. 1627–1637, Dec. 2012.
[26] V. Abolghasemi and A. Ahmadyfard, "An edge-based color-aided method for license plate detection," *Image Vis. Comput.*, vol. 27, no. 8, pp. 1134–1142, Jul. 2009.
[27] A. Prati, I. Mikic, M. M. Trivedi, and R. Cucchiara, "Detecting moving shadows: Algorithms and evaluation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 7, pp. 918–923, Jun. 2003.
[28] H. Asaidi, A. Aarab, and M. Bellouki, "Shadow elimination and vehicles classification approaches in traffic video surveillance context," *J. Vis. Lang. Comput.*, vol. 25, no. 4, pp. 333–345, Aug. 2014.
[29] T. Horprasert, D. Harwood, and L. S. Davis, "A statistical approach for real-time robust background subtraction and shadow detection," in *Proc. IEEE ICCV*, Sep. 1999, pp. 1–19.
[30] K. Mu, F. Hui, X. Zhao, and C. Prehofer, "Multiscale edge fusion for vehicle detection based on difference of Gaussian," *Optik*, vol. 127, no. 11, pp. 4794–4798, Jun. 2016.
[31] J. Kim, J. Baek, and E. Kim, "A novel on-road vehicle detection method using πHOG," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3414–3429, Dec. 2015.
[32] 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.
[33] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, Jun. 2008.
[34] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Sep. 1999, pp. 1150–1157.
[35] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, Sep. 1995.
[36] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.
[37] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
[38] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013.
[39] S. Paisitkriangkrai, C. Shen, and A. V. D. Hengel, "Pedestrian detection with spatially pooled features and structured ensemble learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 6, pp. 1243–1257, Jun. 2016.
[40] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Convolutional channel features," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 82–90.

JIE HUA received the B.E. degree in automation from the Wuhan University of Technology (WHUT), Hubei, China, where he is currently pursuing the M.A. degree in control science and engineering. His research interests include computer vision and deep learning.

YING SHI received the Ph.D. degree in marine engineering from the Wuhan University of Technology (WHUT), Hubei, China, in 2006. She is currently a Professor of artificial intelligence with WHUT. She has published over 40 articles. Her research interests include environment perception technology for safe driving assistance systems and unmanned systems, digital image processing, 3D point cloud data processing, big data analysis, machine learning, and deep learning.

CHANGJUN XIE (Member, IEEE) received the Ph.D. degree in vehicle engineering from WHUT, Wuhan, Hubei, China, in 2009. From 2012 to 2013, he was a Visiting Scholar with UC Davis, Davis, CA, USA. He is currently a Professor with the School of Automation, WHUT. He has published over 50 articles, of which more than 40 are indexed by SCI or EI. His research interests include battery management systems, control strategies of intelligent and connected vehicles, and vehicle control and optimization of new energy vehicles.

HUI ZHANG is currently an Associate Professor with the Intelligent Transportation Systems Research Center (ITSC), WHUT. His main expertise is in the areas of traffic safety management and driving behavior analysis. He has managed more than ten projects related to driving behavior analysis and traffic safety analysis. He has published more than 40 articles. He is a Youth Committee Member of the China Communications and Transportation Association and the Vice Secretary of the Intelligent Transportation Systems Technical Commission of the Chinese Association for Artificial Intelligence. He was awarded the Deborah Freund Paper Award in 2017 by the Transportation Research Board Truck and Bus Safety (ANB70) committee.

JIAN ZHANG received the Ph.D. degree in system optimization and dependability from the University of Technology of Troyes (UTT), Troyes, France, in 2014. He is currently a Lecturer with the School of Automation, WHUT, China. His research interests include reliable and real-time transmission in wireless sensor networks, statistical decision theory, and intelligent information systems.