Smoky Vehicle Detection Based on Improved Vision Transformer

Li Yuan
School of Automation, Southeast University
Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education,
Nanjing, China
mryuanli@foxmail.com

Shuzhen Tong
School of Automation, Southeast University
Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education,
Nanjing, China
114464169@qq.com

Xiaobo Lu*
School of Automation, Southeast University
Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education,
Nanjing, China

ABSTRACT

The harmful exhaust emissions of fuel vehicles in the world are damaging to human health and the environment, thus detecting smoky vehicles from real road environment is significant. At present, methods of smoky vehicle detection based on deep learning have the problem of high false-positive rate. To improve the performance, a two-stage video smoky vehicle detection algorithm based on the smoke classification in the core region from detected vehicle object boxes is proposed in this paper. Specifically, the vehicle object detection is realized by the algorithm based on YOLOv3. The smoke classification is realized by combining Vision Transformer and distillation, and the loss function is optimized in the training process. Experimental results on our smoky vehicle dataset have shown that the improved model achieves an F1 score over 0.4, precision over 0.4, recall nearly 0.1 improvement compared with the basic model, which can effectively reduce the false-positive rate during detection.

CCS CONCEPTS

• **Computing methodologies**; • **Artificial intelligence**; • **Computer vision**; • **Computer vision problems**;

KEYWORDS

Smoke identification, Vehicle detection, Vision transformer, Distillation

ACM Reference Format:
Li Yuan, Shuzhen Tong, and Xiaobo Lu. 2021. Smoky Vehicle Detection Based on Improved Vision Transformer. In *The 5th International Conference on Computer Science and Application Engineering (CSAE 2021), October 19–21, 2021, Sanya, China.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3487075.3487172

1 INTRODUCTION

The sustained and rapid growth of fuel vehicles in the world has brought huge traffic and air pollution problems to major cities in various countries. One of the main sources of air pollution is the emissions of fuel vehicles. Vehicle emissions, which are mainly composed of solid suspended particulates and toxic gases, will have an irreversible impact on human health by contact or inhalation. Haze (PM2.5) and acid rain which pollutes the atmospheric environment and soil environment is also caused by vehicle emissions [1]. Therefore, the detection and alarm of smoky vehicles on the road is of great significance to human health and the ecological environment.

Traditional methods of smoky vehicles detection from outside vehicles mainly include the annual examination of vehicles, the observation by the traffic police on the road and the manual analysis of road surveillance videos. The above methods are time-consuming and arduous. With the increasing amount of road cameras, low-cost and efficient smoky vehicle detection can be achieved by utilizing image processing technology.

Classifying the smoke in images is the key point of smoky vehicle detection. At present, the smoke classification models based on computer vision mainly include traditional machine learning and deep learning methods [2]. The traditional machine learning methods are also mainly divided into two ways: model discriminant and model generation. Appana et al. [3] proposed a model discriminant way for smoke recognition by using the SVM classifier which combines characteristics of the movements, colors and energy features of the smoke. Yuan et al. [4] transformed the task of smoke classification into Gaussian process regression of data and parameters. The above all traditional methods for smoke recognition can only combine minority features. Massive calculations will be generated if models combine with more features. The application scene is narrowed due to the poor adaptability of the algorithm.

Deep learning can extract features automatically and solve the above problems in traditional smoke classification methods. Tao et al. [5] presented a three-stage framework for smoky vehicle detection to improve the algorithm performance. Zhang et al. [6] purposed a smoky vehicle detection method by introducing transfer learning to MobileNet, VGG19 and other CNN classification networks, but the high false-positive rate still cannot be solved. Cao et al. [7] combined spatial-temporal information with Inception V3 and LSTM to achieve a better result. Recently, Transformers [8] have attracted more attention in computer vision. Good performances have shown in many specific tasks [9]. One of the representative

[*] Corresponding author: xblu2013@126.com

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CSAE 2021, October 19–21, 2021, Sanya, China
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8985-3/21/10…$15.00
https://doi.org/10.1145/3487075.3487172

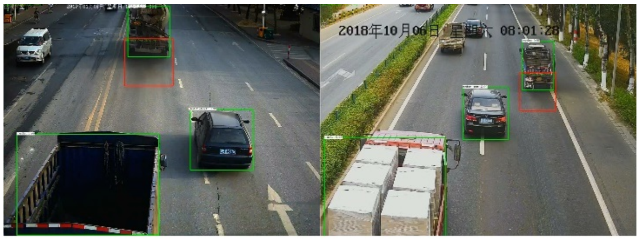

Figure 1: Examples of Smoky Vehicle Detection from Actual Road Videos.

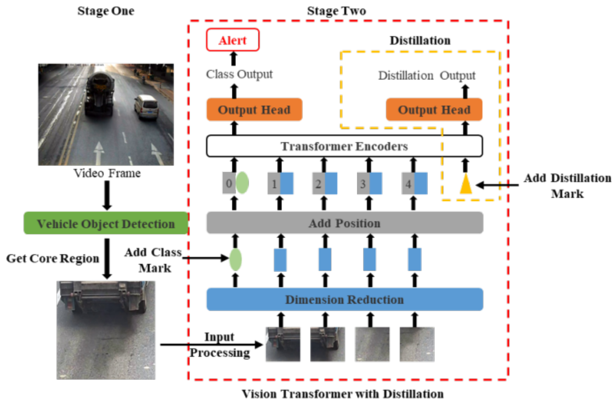

Figure 2: Smoky Vehicle Detection Algorithm Overview, the Classifier in the Red Box is the Main Part of it.

works in the image classification task is the Vision Transformer (ViT) [10], which is an entire attention-based model without using CNNs.

The purpose of this paper is to realize an effective and automatic algorithm for detecting smoky vehicles from video frames which are obtained by the actual road environment camera under the condition of unmanned surveillance. After getting the tail core region of detected vehicle boxes, we purpose a method of combining ViT model with knowledge distillation [11] and optimizing the training process to decrease the high false-positive rate in smoky vehicle detection. The result of experiments shows that an F1 score over 0.4, precision over 0.4, recall nearly 0.1 is obtained by our method.

2 METHOD

In this section, the vehicle object detection and smoke classification models used for smoky vehicle detection are described in detail. The example of real-time smoky vehicle detection in the road surveillance video is shown in Figure 1, where the smoke emitted by smoky trucks is in the red box. Smoke images are saved by the algorithm after detection.

As shown in Figure 2, the real-time smoky vehicle detection process for video frames from actual road videos is mainly a two-stage process. First, we detect vehicle objects from each video frame, and then expand the tail region of the detected vehicle object box as the core region of the smoke classifier. Second, the smoke classifier is used to classify the smoke in the core region according to the characteristics of the vehicle smoke. This two-stage algorithm is purposed to improve the speed and accuracy of smoke recognition. Finally, if smoke is existing for several consecutive frames, the alarm will be given. An overview of the algorithm is depicted in Figure 2

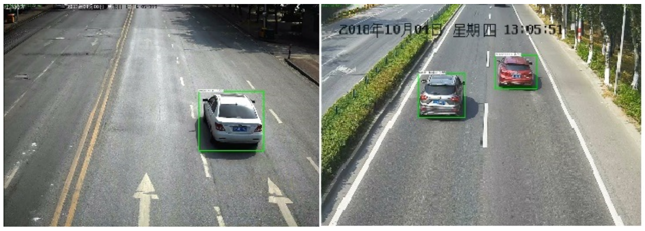

Figure 3: Examples of Vehicle Object Detection Realized by YOLOv3.

2.1 Vehicle Object Detection Algorithm

Classify the smoke directly from the whole image is unrealistic because there are shadows and other interference in the actual environment, which may greatly increase the false-positive rate. Therefore, classifying smoke from the tail region of vehicles especially the area where the exhaust pipe located is a more effective way. In this paper, vehicle object detection is realized by using YOLOv3 [12] algorithm, which is a classic work of You Only Look Once series. YOLOv3 is one of the representative works in the one-stage detection network which doing regression from images directly. Under the COCO mAP-50, its detection speed is about four times over other algorithms in the same period. After training the network with labeled images from the actual road environment, the model can effectively detect vehicle targets and satisfy the demands of the algorithm. Examples of vehicle object detection results realized by YoLov3 are shown in Figure 3

The core region for smoke classification can be obtained by scaling the tail region from the vehicle object box. As shown in Figure 4, in order to get the tail region, we produce a square whose side equals the width of the vehicle object box at 70% of the height of the vehicle object box. To avoid missing caused by the diffusion of the smoke, the size of the core region is enlarged from the tail region square. Core regions are inputs sent to the classifier.

2.2 Smoke Classify Network

In this paper, the smoke is classified by ViT of Transformer architecture. Meanwhile, we combine distillation and optimization of the loss function to achieve better model training results.

2.2.1 Basic ViT Network. The idea of basic ViT is to keep the main part of Transformer which is mainly used in NLP field unchanged, and then adapt the input images to the model by using the original encoder to make predictions of images.

Image input processing: To satisfy the input requirement of original NLP tasks, images of (H, W, C) can be cut into m pieces with the size of $(P{\times}P{\times}C)$ patches, and then patches of $m{\times}(P, P, C)$ can be extended into m sequences of length with $P^2 \times C$. Due to the length

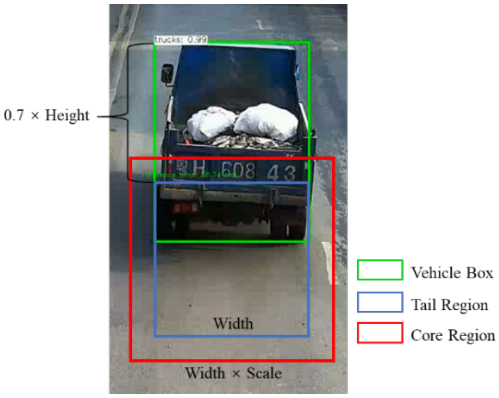

Figure 4: Construction of the Core Region for Smoke Classification.

of sequences obtained directly is large, the fully connected layer being used to reduce the dimension of sequences in the algorithm.

Adding a class mark and position information: ViT abandons the structure of the decoder in Transformer, thus the model adds a trainable class mark which is responsible for representing the classification prediction results at the front of input sequences. Then, a position code is added for avoiding information confusion caused by the weakness of the Attention mechanism within Transformer.

Transformer Encoder: The encoder structure mainly includes two processes: Muti-head self-Attention and feed-forward network. In this paper, Muti-head self-Attention combines eight self-Attention modules. Each self-Attention module carries on the operation:

$$Z_i = soft\max(QK^T/d)V \tag{1}$$

Where Q, K, V are Query vector, Key vector and Value vector. They can be obtained by multiplying the encoder input and their own trainable matrixes. d is the normalized factor. The output of Muti-head self-Attention is obtained by a fully connected layer after weighted combination of eight Z_i. This model replaces ReLU with GELU to achieve better results in the feedforward network., Residual channels and layer normalization are added to the encoder in order to improve the speed of optimization and avoid degradation.

Finally, after layer normalization and a fully connected layer, the output predicted value of the model for different classes can be reached by extracting the class mark.

2.2.2 Distillation. A huge data set is required by ViT to fully train the Transformer, which is unbearable for individuals. Distillation is to solve this problem by adding an additional distillation mark at the end of the input sequence, just as adding a class mark. We introduce a pre-trained Resnet-50 model as an adviser model, which makes the distillation mark output close to the adviser model output. During the back propagation learning process, both the original class mark and distillation mark are trained at the same time, which can optimize the ViT student model better in the interaction with the adviser model.

2.2.3 Training Process. The training process of this model is using Adam algorithm to back propagate the error of the calculated gradient and then updating parameters. In original ViT, the algorithm uses the cross-entropy error loss function which combined with Log-Softmax and negative log-likelihood loss (Eq. 2).

$$L_{CE} = \log \sum_{j=0}^{N-1} e^{output(j)} - \log e^{output(label)} \tag{2}$$

The distillation process integrates class mark and distillation mark (Eq. 3). Let Z_s be the output of the student model, Z_a the output of the adviser model. We denote by k the coefficient of distillation, α the equilibrium coefficient, S the softmax function. L_{KL} is the Kullback-Leibler divergence loss (Eq. 4).

$$L_D = (1-\alpha)L_{CE}(S(Z_s), label) + \alpha k^2 L_{KL}(S(Z_s/k)), S(Z_a/k)) \tag{3}$$

$$L_{KL} = \frac{1}{N}\sum_{n=0}^{N}(\log y_n - x_n)y_n \tag{4}$$

In the next section, in order to achieve better training performance, we attempt to replace the cross-entropy loss function with binary cross-entropy with logits loss (Eq. 5) or smooth L1 loss (Eq. 6):

$$L_{BCE} = -\frac{1}{N}\sum_{n=0}^{N}(y_n \log \sigma(x_n) + (1-y_n)\log(1-\sigma(x_n))) \tag{5}$$

$$L_{smoothL1} = \frac{1}{N}\sum_{n=0}^{N} z_i, z_i = \begin{Bmatrix} 0.5(x_i - y_i)^2 & |x_i - y_i| < 1 \\ |x_i - y_i| - 0.5 & |x_i - y_i| \geq 1 \end{Bmatrix} \tag{6}$$

where$N$is the batch size, $\sigma$is the sigmoid function, $x_n$is the output of the model, $y_n$is the label value.

3 EXPERIMENTS

In order to achieve two-stage smoky vehicle detection, the dataset in this paper consists of two parts: vehicle object detection dataset and smoke classification dataset. In vehicle object detection, the model could be trained by transfer learning, but test results were not satisfactory due to the tiny dataset. We utilized data augmentation to train the vehicle object detection model directly.

The smoke classification dataset was labeled by two parts, positive and negative, in which smoke existing images were divided into positive samples. Square images were used for training and testing in order to satisfy the input of the classifier. Because the proportion of smoky vehicles is small in the actual road environment, data augmentation of brightness changing and horizontal flipping were used in the training dataset to enhance the robustness of the model for feature extraction. The training dataset contains 34016 images, of which includes 6107 positive samples. In order to be close to the application of the actual road environment, the testing dataset came from video frames captured by a road camera during the day. We then manually extracted the tail area of vehicles in images. A total of 18598 testing images were obtained by this method, of which 27 were positive samples including 11 difficult samples. Examples of smoke classification dataset are shown in Figure 5

In this section, the evaluation of the vehicle object detection model is mAP (Eq. 7), which balances the detection accuracy of each category. For the smoke classification model, analysis based on accuracy is invalid, because the proportion of smoky vehicles in the actual road environment is very small. Under extremely unbalanced testing conditions, if the model does not have the ability to distinguish smoke at all, the accuracy will still be over 99%.

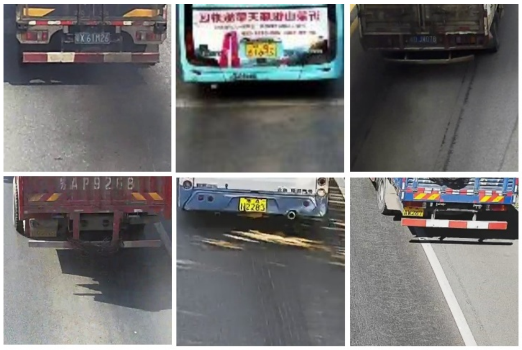

Figure 5: Examples of Smoke Classification Dataset, the Positive Samples are in the First Row and the Negative Samples are in the Second Row.

Table 1: Results of Vehicle Object Detection Experiments

Model	mAP	Time	Size
YOLOv3	85.4%	35ms	492.8MB
YOLOv3-tiny	78.2%	12ms	69.5MB
YOLOv3-SPP	73.9%	28ms	501.2MB
YOLOv3-Transfer	56.3%	37ms	246.8MB

Therefore, with comprehensive consideration of precision, recall and F1 score (Eq. 8) is the main evaluation index in the smoke classification model.

$$mAP = \sum_{class} precision_{class}/class \tag{7}$$

$$F1score = 2 \times \frac{precision \times recall}{precision + recall} \tag{8}$$

After training the vehicle object detection model on the small dataset obtained from the actual road, taking mAP as the main evaluation index, we chose the YOLOv3-tiny model which got high accuracy and the shortest detection time to detect the vehicle target. It fits the needs of real-time detection and computational requirements of practical engineering. As shown in Table 1, when the dataset is too small, the actual results of transfer learning and spatial pyramid pooling [13] were unacceptable, so we did not choose to add them.

3.1 Training Effect

The smoke classification network is the core of smoky vehicle detection in this paper. In order to better improve the performance of the smoke classification model, this section mainly experiments on two aspects. Experiment 1 mainly studies the effects of loss functions on the basic classification model. Experiment 2 mainly compares performances based on adding distillation. All models in Experiment 1 and Experiment 2 are trained by iterating 20 epochs with exactly the same dataset.

We replaced the cross-entropy (CE) loss function from the basic network with smooth L1 loss (smooth) or binary cross-entropy with logits loss (BCE with logits) to study the effects of different loss functions in the model training process. The results of Experiment 1 are shown in Table 2

Smooth L1 loss uses softer gradient values. It can avoid exploding gradients caused by excessive error. But the results of it cannot reach the BCE with logits loss in the experiment.

3.2 Influence of Distillation

Considering the effective role of BCE with logits loss in the training process of the smoke classification model based on the original ViT in Experiment 1, we replaced the loss function of the distillation model for calculating the loss value between model outputs and the labels within the student model. The results of Experiment 2 are shown in Table 3

As shown in Table 3, distillation can effectively reduce the false-positive rate which the model mistakenly predicts a non-smoke image into a smoke one. It effectively improved the classification ability of the model by balancing precision and recall. By introducing the distillation mark, an additional pre-trained Resnet-50 model can guide the data mark during the training process. This is equivalent to strengthen the supervision role of the model training, thus reducing the training cost and improving the ability of the model. Through Experiment 1, we can introduce the promotion impact of BCE with logits loss in the training process. F1 score over 0.4, precision over 0.4, recall nearly 0.1 can be obtained by our method that combined distillation and optimizing training loss from the original ViT model.

4 CONCLUSION

This paper purposed a complete framework for detecting smoky vehicles from videos. By extracting the core region from vehicle objects and then sending it to the classifier, the smoky vehicle targets in the video can be detected effectively. In vehicle object detection, we chose the YOLOv3-tiny model which got a high accuracy and the shortest detection time to detect the vehicle target. Based on the classification model ViT, we studied and analyzed the effects of distillation with an additional distillation mark and the impacts of different loss functions on the training process. The final results showed that the improved ViT model which combined with distillation and BCE with logits loss improved F1 score over 0.4, precision over 0.4, recall nearly 0.1 compared with the original model. The optimized model can effectively reduce the false-positive rate. Actually, the model still has a weakness to identify difficult samples, and the false-positive rate still needs to be improved. Therefore, smoke classification of video sequences rather than single video frames is an important direction of future work.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grant 61871123.

Table 2: Comparison of Experimental Results of Different Loss Functions

Model	Accuracy	Precision	Recall	F1 score
ViT-base	0.9842	0.0412	0.4444	0.0755
ViT-smooth	0.9933	0.1603	0.5833	0.2515
ViT-BCE	0.9952	0.2453	0.7222	0.3662

Table 3: Comparison of Experimental Results of Improved ViT Model

Model	Accuracy	Precision	Recall	F1 score
ViT-base	0.9842	0.0412	0.4444	0.0755
ViT-distillation	0.9975	0.2432	0.3333	0.2812
ViT-ours	0.9979	0.4524	0.5278	0.4872

REFERENCES

[1] Y. Ni (2020). Vehicle Exhaust Emission Problem and Energy Saving and Emission Reduction Method. Technology & Management, (10),52-53.

[2] T. Zhang, W. Yang, J. Zhang and K. Peng, (2021). Video black smoke detection methods for vehicles: a survey. Journal of Image and Graphics, 26(02), 316-333.

[3] D. K. Appana, R. Islam, S. A. Khan and J. M. Kim (2017). A video-based smoke detection using smoke flow pattern and spatial-temporal energy analyses for alarm systems. Information Sciences, 418, 91-101.

[4] F. Yuan, X. Xia, J. Shi, H. Li and G. Li (2017). Non-linear dimensionality reduction and Gaussian process based classification method for smoke detection. IEEE Access, 5, 6833-6841.

[5] Tao, H., Zheng, P., Xie, C., and Lu, X. (2020). A three-stage framework for smoky vehicle detection in traffic surveillance videos. Information Sciences, 522, 17-34.

[6] Zhang, G., Zhang, D., Lu, X., & Cao, Y. (2019, November). Smoky Vehicle Detection Algorithm Based On Improved Transfer Learning. In 2019 6th International Conference on Systems and Informatics (ICSAI) (pp. 155-159). IEEE.

[7] Cao, Y., & Lu, X. (2019). Learning spatial-temporal representation for smoke vehicle detection. Multimedia Tools and Applications, 78(19), 27871-27889.

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762.

[9] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2021). Transformers in Vision: A Survey. arXiv preprint arXiv:2101.01169.

[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[11] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2020). Training data-efficient image transformers & distillation through attention. arXiv preprint arXiv:2012.12877.

[12] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.

[13] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 37(9), 1904-1916.FNM Surname (2018). Article Title. Journal Title, 10(3), 1–10.