浙江科技学院学报，第32卷第3期，2020年6月
Journal of Zhejiang University of Science and Technology
Vol. 32 No. 3, Jun. 2020

doi：10.3969/j.issn.1671-8798.2020.03.007

基于图像分割的驾驶员分心行为识别研究

叶　绿[a]，陈　铖[b]，Sugianto Sugianto[a]，Chido Natasha Muponda[a]，
Agordzo George Kofi[a]，Koi David Ernest[a]
（浙江科技学院 a.信息与电子工程学院；b.机械与能源工程学院，杭州 310023）

摘　要：针对驾驶员分心行为对交通安全产生隐患的问题，提出一种结合图像分割与卷积神经网络的驾驶员分心行为检测方法。该方法通过使用图像分割处理后的驾驶员不同分心行为的图像对卷积网络结构模型进行训练，来减轻背景噪声的影响，以提高模型的识别性能。试验中使用未经分割的图像与经过前景分割后图像分别训练卷积神经网络模型，用分割后图像训练的模型识别的准确率达到了93.84%，高于使用原图像训练的模型。试验结果表明，结合图像分割和深度学习的驾驶员分心行为检测方法对驾驶员的分心行为有较好的检测效果。

关键词：图像分割；神经网络；特征提取；分心行为识别
中图分类号：TP391.41　　**文献标志码**：A　　**文章编号**：1671-8798(2020)03-0209-07

Driver's distracted behavior recognition research based on image segmentation

YE Lü[a], CHEN Cheng[b], Sugianto Sugianto[a], Chido Natasha Muponda[a],
Agordzo George Kofi[a], Koi David Ernest[a]
(a. School of Information and Electronic Engineering; b. School of Mechanical and Energy Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, Zhejiang, China)

Abstract: In response to the problem of traffic safety caused by driver's distracted behavior, a detection scheme of driver's distracted behavior is proposed in combination with image segmentation and convolutional neural network. This method used driver's different distracted behavior images processed by image segmentation to train the convolutional neural network structure model, to reduce effect of background noise and improve accuracy of the model. The experiment used the unsegmented images and the foreground segmented images to train the convolutional neural network model respectively, with recognition accuracy rate of the model trained by segmented image up to 93.84%, greater than that of the model trained by unsegmented

收稿日期：2019-10-28

通信作者：叶　绿（1962—　），女，浙江省杭州人，教授，博士，主要从事人工智能和计算机视觉研究。E-mail：2607523678@qq.com。

image. The experimental results show that the driver's distracted behavior detection method combined with image segmentation and deep learning can detect the driver's distracted behavior greatly.

Keywords: image segmentation; neural network; feature extraction; distracted behavior recognition

社会经济的发展离不开交通的发展,交通安全是全世界关注的问题,而驾驶员的状态和行为则是影响交通安全的主要因素。随着车内多媒体设备的多样化,人们越来越倾向于在车内进行多件事情的处理。有些人在驾驶过程中进行除驾驶以外的动作,导致注意力无法完全集中在驾驶上,由此产生交通安全的隐患。有研究表明驾驶过程中因打电话或发短信而发生事故的概率分别增加了3倍和4倍[1]。因此,有必要对分心驾驶行为进行研究以减少其带来的交通隐患[2]。

驾驶员的状态与行为已引起国内外研究者的关注,研究大致分为使用传统的机器学习算法和深度学习算法这两种。对于传统的机器学习算法,目前相关的研究大多通过自己设定的特征,并基于这些特征使用传统算法来分辨驾驶员的分心行为。Kutila 等[3]通过获取驾驶员的视觉、头部和车道保持差异的数据,使用支持向量机(support vector machine,SVM)方法来实现对驾驶员分心行为的检测与分类,其识别准确率在65%~80%。Sahayadhas 等[4]提出一种基于生理信息的驾驶员状态的检测方法,通过对驾驶员心电图、肌电图等实测得到的生理信息,再借用 k 邻近和线性判别方式对信息所包含的特征进行分析,结果表明心电信号和表面肌电信号是测量驾驶员注意力的一个重要指标。杨晓峰等[5]通过使用ASM算法(active shape model,主动形状模型)来提取特征点而得到驾驶员脸部特征点位置信息,再使用支持向量机对该信息进行分类来检测驾驶员的低头行为,其识别准确率达94%。Ebadi 等[6]的试验结果表明,当驾驶员在免提手机的情况下进行驾驶时,眼球对周围场景的扫视比例显著降低,不容易发现驾驶环境中的潜在危险。在上述部分方法中,对驾驶员生理信息的测量及眼球的监控都需要特殊的硬件设备。

随着深度学习方法的兴起,许多研究者尝试将其应用到驾驶员行为的检测中。而这类方法大多基于计算机视觉,只需获取驾驶员的图像信息,通过深度学习算法自动提取图像特征。Le 等[7]尝试使用多尺度快速 RCNN(region convolutional neural networks,区域卷积神经网络)方法来检测驾驶员是否使用手机、是否单手驾驶或双手离开方向盘,通过这些特征来判断驾驶员的状态。杨星等[8]借用体感相机采集驾驶员行为的深度图像信息,加上关节位置信息,利用随机森林和最大信息系数法得出不同特征的重要性,再借助前馈神经网络进行驾驶员不同行为的识别。Masood 和 Celaya-Paddilla 等[9-10]均使用了卷积神经网络来检测驾驶员的分心行为,前者定义了1种正常驾驶状态及9种分心行为,后者则通过车内安装广角摄像头来获取数据,从而进行驾驶员发短信行为与正常驾驶状态的识别。胡军等[11]提出了一个轻量级的半级联网络结构,该网络通过提取人脸和手部的形态学特征来对驾驶员的行为做出判断。

传统识别方法存在特征提取成本高、识别性能或范围有限的问题。相比较而言,深度学习方法具有一定的优势。基于此,本研究提出一种使用图像分割去除背景冗余信息来提升卷积神经网络识别驾驶员分心行为能力的方法。

1 图像分割的预处理

在将图像数据输入神经网络模型之前,使用 opencv 中定义实现的 GrabCut 算法对图像进行前后景分割,以减少网络的训练时间及图像背景的影响,达到提高网络分类准确性的目的。

GrabCut 算法[12]在图割算法的基础上使用了高斯混合模型(Gaussian mixture model,GMM)代替原本的灰度直方图模型来对图像中的目标和背景进行建模,并将模型的能量最小化方案改为迭代算法来优化模型中的参数。算法实现的简要流程如下:首先在图像中定义包含前景对象的矩形,该矩形外的图像被视作背景,矩形内包含的图像像素则被定义为可能是前景的像素;然后通过这些已经定义的像素使用

GMM 对前景和背景部分进行建模；在初步建模后给图像中的每个像素分配高斯分量，然后基于已经归类的像素来学习优化 GMM 的参数，再借助整个图像的能量式建立图后进行分割估计；迭代初步建模后的 3 个步骤，直到图像的能量式收敛。

设 α 为图像中像素的标签（背景 0 或前景 1），$\boldsymbol{k}$ 为 GMM 的分量，θ 为 GMM 的参数构成的集合，$\boldsymbol{z}$ 为该像素的 RGB 值，整个图像的能量式为

$$E(\alpha,\boldsymbol{k},\theta,\boldsymbol{z}) = U(\alpha,\boldsymbol{k},\theta,\boldsymbol{z}) + V(\alpha,\boldsymbol{z})。\tag{1}$$

式(1)中的区域项

$$U(\alpha,\boldsymbol{k},\theta,\boldsymbol{z}) = \sum_n D(\alpha_n,\boldsymbol{k}_n,\theta,\boldsymbol{z}_n)。\tag{2}$$

式(2)中：

$$D(\alpha_n,\boldsymbol{k}_n,\theta,\boldsymbol{z}_n) = -\lg\pi(\alpha_n,\boldsymbol{k}_n) + \frac{1}{2}\lg\det\boldsymbol{\Sigma}(\alpha_n,\boldsymbol{k}_n) +$$

$$\frac{1}{2}[\boldsymbol{z}_n - \boldsymbol{\mu}(\alpha_n,\boldsymbol{k}_n)]^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\alpha_n,\boldsymbol{k}_n)[\boldsymbol{z}_n - \boldsymbol{\mu}(\alpha_n,\boldsymbol{k}_n)]。$$

由于 GMM 的参数分别为高斯分量权重 π、高斯分量均值向量 $\boldsymbol{\mu}$、协方差矩阵 $\boldsymbol{\Sigma}$，那么模型参数

$$\theta = \{\pi(\alpha,\boldsymbol{k}),\boldsymbol{\mu}(\alpha,\boldsymbol{k}),\boldsymbol{\Sigma}(\alpha,\boldsymbol{k}),\alpha = 0,1,\boldsymbol{k} = \boldsymbol{k}_1,\cdots,\boldsymbol{k}_n\}。$$

式(1)中的边界项

$$V(\alpha,\boldsymbol{z}) = \gamma\sum_{(m,n)\in C}\exp(-\beta\|\boldsymbol{z}_m - \boldsymbol{z}_n\|^2);\tag{3}$$

$$\beta = (2\langle(\boldsymbol{z}_m - \boldsymbol{z}_n)^2\rangle)^{-1}。$$

式(3)中：C 为相邻像素对的集合，该集合中像素对的标签 $\alpha_n \neq \alpha_m$。

完成建模后对该模型进行迭代最小化，分为以下 3 个步骤：

第一步，将 GMM 分量分配给给定图像的像素

$$\boldsymbol{k}_n := \underset{\boldsymbol{k}_n}{\arg\min} D_n(\alpha_n,\boldsymbol{k}_n,\theta,\boldsymbol{z}_n)。$$

第二步，从给定的图像中得到 GMM 参数

$$\theta := \underset{\theta}{\arg\min} U(\alpha,\boldsymbol{k},\theta,\boldsymbol{z})。$$

第三步，进行分割估计

$$\min_{\{\alpha_n:n\in T_U\}}\min_{\boldsymbol{k}} E(\alpha,\boldsymbol{k},\theta,\boldsymbol{z})。\tag{4}$$

式(4)中：T_U 为可能属于前景的像素集合。

重复上述 3 个步骤直到能量式收敛。GrabCut 算法的图像分割效果如图 1 所示。

(a) 原始数据　　(b) 矩形框定　　(c) 背景分割

图 1　GrabCut 算法的图像分割效果

Fig. 1　Image segmentation of GrabCut algorithm

2　卷积神经网络的运用

深度学习算法选择了在计算机视觉领域表现优秀的卷积神经网络（convolutional neural networks，CNN）。卷积神经网络的概念[13]是 LeCun 提出的，其提出的 LeNet5 模型是最早的卷积神经网络模型，该网络架构的出现使得图像处理领域取得了突破性的进展。传统的卷积神经网络模型由特征提取和数

据分类两部分组成。特征提取主要通过卷积层来实现。通过特定的网络结构设计提取满足算法需求的图像特征后，分类模块会根据提取出来的特征对目标图像进行分类。卷积神经网络由多个卷积层组成，这些卷积层能通过使用不同大小的卷积核提取图像的识别特征，在经过多个卷积层提取到图像的高级特征后，分类层再借助这些特征给出图像的类别。卷积神经网络主要通过增加隐藏层的深度来降低图像的维数，从而使模型能够在低维空间中提取到稀疏图像的特征。

3 试验设计与分析

3.1 试验流程

本试验主要由两部分组成：第一部分对数据集中的图像使用 GrabCut 算法进行图像分割，保存前景部分，并将处理后的图像进行训练集与测试集的划分；第二部分是先读取训练集的图像，对其进行数据增强和预处理，然后将处理过的图像输入到 CNN 分类模型当中来训练模型的参数，最后使用测试集验证训练得到的模型来获得识别结果。训练及测试流程如图 2 所示。

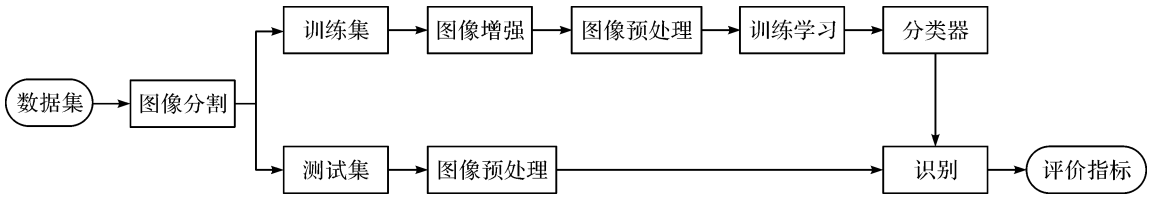

图 2 训练及测试流程

Fig. 2 Process of training and test

3.2 数据集的建立

试验使用的数据集是 StateFarm 提供的公开数据集，我们在该数据集基础上加入自制的驾驶员行为数据集来增强数据的差异性，以提升训练模型的泛化能力。新数据集的建立过程如图 3 所示。

图 3 新数据集的建立过程

Fig. 3 Establishment process of new dataset

选择 StateFarm 数据集中的 26 名司机及自制数据集的 5 名司机的 5 类不同行为图像：专心驾驶、打字、打电话、喝水、转身取物。由于该数据集的图像均为视频中截取的视频帧，会存在相同驾驶员高度相似的图像。若训练集与测试集存在高度相似图像则会影响试验的准确性，因此需选择不同司机的图像作为训练集及测试集。本试验选择 26 名司机的 11 400 张图像作为训练集，剩余 5 名司机的 2 192 张图像作为测试集。进行试验时，先对所有的图像进行分割处理，以减少背景的影响，分割处理后的数据集图像如图 4 所示。

(a) 专心驾驶 (b) 打字 (c) 打电话 (d) 喝水 (e) 转身取物

图 4 分割处理后的数据集图像

Fig. 4 Segmented dataset images

3.3 数据的增强与预处理

首先将分割后 640×480×3 的 RGB 图像调整为 224×224×3 的图像以符合网络模型的输入要求，然后对训练集的图像进行水平变换，旋转，最大裁剪生成一组新的图像来达到数据增强的目的。数据增强可以提供一个更为丰富的数据集来训练增强分类器的性能。在图像输入网络模型之前还需要对所有图像进行预处理。从每个像素值中减去 RGB 通道在所有像素上的平均值，减去平均值之后可以使数据

值居中，然后使用梯度反向传播来训练模型。另外，每个特征都有一个相似的范围，这可以防止梯度超出特定范围。CNN 的一个特点是参数共享，如果没有将输入缩放到相似范围的值，则共享并不容易实现。

3.4　识别模型的选取

VGG16[14]（Visual Geometry Group，牛津大学计算机视觉组）网络模型的输入是 RGB 图像。图像输入后会通过一系列的卷积层，这些卷积层使用感受野非常小的 3×3 的卷积核，卷积的步长固定为 1。对于 3×3 的卷积核，其空间填充也设为 1，该模型使用 5 个最大池化层，这些池化层置于前部分卷积层后面。池化窗口大小设为 2×2，步长为 2。网络模型的最后部分是 3 个全连接层，最后一层的激活函数使用归一化指数函数来给出分类结果。VGG16 模型结构如图 5 所示，conv3-64 代表 64 个 3×3 卷积核的卷积层，其余同理。

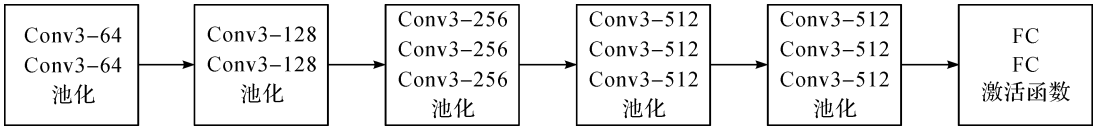

图 5　VGG16 模型结构

Fig. 5　VGG16 model structure

InceptionV3[15]在原有 Inception 结构上将大尺度的卷积分解成多个小尺度卷积，并且加入了非对称卷积，将部分 3×3 的卷积核分解成 1×3 和 3×1 的卷积核，大大减少了模型的计算量。InceptionV3 模块结构如图 6 所示。

Xception[16]是在 InceptionV3 上做出的一种改进，主要借鉴深度可分离卷积的操作将通道卷积与空间卷积进行分类，并在网络中加入了残差连接的结构。

MobileNetV2[17]与 Xception 同样应用了深度可分离卷积及残差连接，并在此基础上做出了修改。该网络模型去掉了深度可分离卷积模块后面的 ReLU6函数，使得其中的残差连接变成了与传统残差结构相反的“逆向残差连接”。

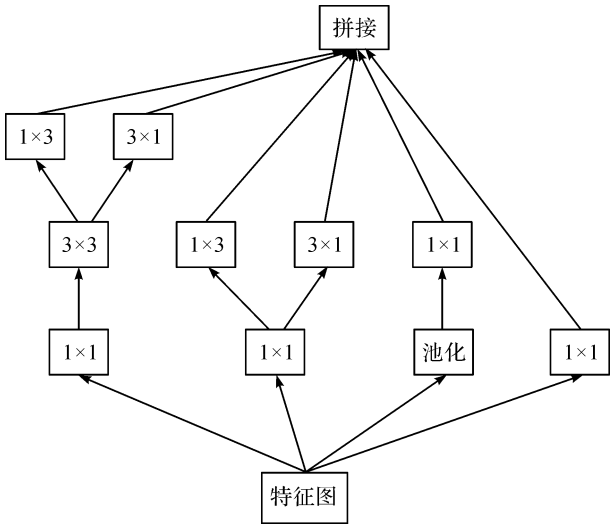

图 6　InceptionV3 模块结构

Fig. 6　Module structure of inceptionV3

3.5　试验与结果分析

试验运行环境为操作系统 win10 64 位，GPU GTX1060 6 GB，电脑内存 16 GB。仿真基于 python 语言实现，网络模型通过 keras 和 tensorflow 框架搭建。

首先将数据集中的所有图像路径读到内存中，调用 opencv 实现的 GrabCut 算法，定义函数相关参数，对所有图像进行前后背景分割后保存为分割后的数据集，同时记录该算法运行的时间。分割一张图像的时间约为 0.163 6 s。然后重新将划分好的训练集与测试集的图像和对应的标签读入到程序中。使用 keras 搭建 VGG、Inception 和 Xception 模型结构，定义优化方式为随机梯度下降，随机梯度下降的动量为 0.9，学习率初始化为 1×10^{-3}，衰减率为 1×10^{-6}；损失函数使用交叉熵函数、以损失函数的值和全局准确性为评价标准。编译模型，在输入数据到模型之前先将图像调整为 224×224×3 的图像，经过预处理和数据增强后，和标签一起输入模型中训练模型。每一批量的训练样本数为 12，使用卷积网络提取特征，训练 50 个迭代，若测试集损失函数超过 10 个迭代未下降，则训练提前停止，保存在测试集上损失函数最优的模型，并记录该模型识别结果的准确率（正确识别的图像数量/所有图像数量）。准确率计算方式如下：

$$T_{acc}=\frac{TP+TN}{TP+TN+FP+FN}。$$

再次使用测试集输入该模型输出识别结果的混淆矩阵，观察每一行为类别的识别效果，并记录所消耗的时间。使用原始图像和分割后的图像分别训练相同结构的网络模型，再使用 3 个不同结构的网络模型进行对比试验，避免试验结果的偶然性及验证更好的网络结构可以提升识别的性能。共进行 6 次试验，结果对比见表 1。

表 1 试验结果对比

Table 1 Comparison of experimental results

网络模型结构	图像分割	准确率/%	测试耗时/s
VGG16	否	62.36	19.56
	是	85.08	19.14
InceptionV3	否	81.42	17.96
	是	90.28	17.27
Xception	否	82.92	17.73
	是	93.84	16.83
MobileNetV2	否	87.16	12.88
	是	92.95	12.47

由表 1 可知，对同一种网络结构，使用分割后的图像作为数据集训练的模型比使用原图像作为数据集训练的模型拥有更优秀的识别准确性，VGG16 模型性能提升了 20%左右，InceptionV3 和 Xception 的识别效果均提升了 10%，MobileNetV2模型性能提升了约5%。其主要原因是分割后的图像去除了背景中包含的冗余信息而保留了人体行为的关键信息，大大减少了噪声的影响，使神经网络能够侧重学习人体动作的特征，分类器达到更高的准确率。同时由于图像所包含信息量的减少，模型对分割后图像的识别速度也得到了部分提升。但是相比较模型的识别速度的略微提升，图像分割操作需花费较长时间，并不能很好地满足驾驶员行为识别的实时性要求。另外，通过该方法训练神经网络分类器可以达到识别不同驾驶员行为的目的，除了使用原始图像训练的 VGG16 外，其他试验的识别结果均达到 80%以上，结合图像分割最好的 Xception 模型性能达到 93%。图 7 是使用分割图像训练的 Xception 模型在测试集上识别结果的混淆矩阵，图中的对角线为识别正确的图像的数量。从图 7 中发现，模型对“专心驾驶”及“打字”状态的识别还存在一定的误差，对剩余 3 种状态的识别表现良好。

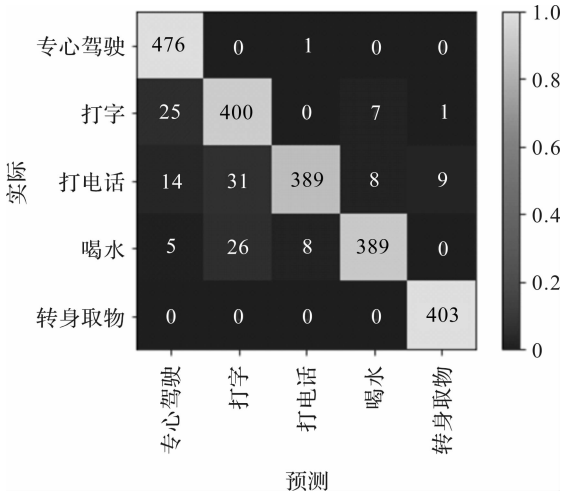

图 7 使用分割数据的 Xception 的混淆矩阵

Fig. 7 Confusion matrix of Xception with segmented data

4 结 语

本文提出一种准确率较高的针对驾驶员分心行为的检测方法。该方法主要通过对数据集进行图像分割预处理来提高神经网络分类器的性能。相比较传统的神经网络算法，结合图像分割的神经网络算法对驾驶员的分心行为具有更优的识别性能，但是识别速度受一定的限制。在后续研究中可基于此开发完整的驾驶员分心行为警示系统，使用速度更快的图像分割算法以满足驾驶员行为识别的实时性要求。

参考文献：

[1] CHOUDHARY P, VELAGA N R. Mobile phone use during driving: effects on speed and effectiveness of driver compensatory behaviour[J]. Accident Analysis & Prevention, 2017, 106(6): 370.

[2] 文鑫垚.关于分心驾驶行为的综述[J].青海交通科技，2019(2)：24.

[3] KUTILA M H，JOKELA M，MÄKINEN T，et al. Driver cognitive distraction detection：feature estimation and implementation[J]. Journal of Automobile Engineering，2007，221(9)：1027.

[4] SAHAYADHAS A，SUNDARAJ K，MURUGAPPAN M，et al. A physiological measures-based method for detecting inattention in drivers using machine learning approach[J]. Biocybernetics and Biomedical Engineering，2015，35(3)：198.

[5] 杨晓峰，邓红霞，李海芳.基于计算机视觉的驾驶员低头行为检测[J].Computer Science，2016，43(增刊 1)：210.

[6] EBADI Y，FISHER D L，ROBERTS S C. Impact of cognitive distractions on drivers' hazard anticipation behavior in complex scenarios[J]. Transportation Research Record：Journal of the Transportation Research Board，2019，2673(9)：440.

[7] LE T H N，ZHENG Y，ZHU C，et al. Multiple scale faster-RCNN approach to driver's cell-phone usage and hands on steering wheel detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Las Vegas：IEEE，2016：46.

[8] XING Y，LV C，ZHANG Z，et al. Identification and analysis of driver postures for in-vehicle driving activities and secondary tasks recognition[J]. IEEE Transactions on Computational Social Systems，2018，5(1)：95.

[9] MASOOD S，RAI A，AGGARWAL A，et al. Detecting distraction of drivers using convolutional neural network [EB/OL]. (2018-01-12)[2019-09-08]. http://www.sciencedirect.com/science/article/pii/S0167865517304695.

[10] CELAYA-PADILLA J M，GALVÁN-TEJADA C E，LOZANO-AGUILAR J S A，et al. "Texting & driving" detection using deep convolutional neural networks[J]. Applied Sciences，2019，9(15)：2962.

[11] HU J，LIU W，KANG J，et al. Semi-cascade network for driver's distraction recognition[J]. Proceedings of the Institution of Mechanical Engineers，Part D：Journal of Automobile Engineering，2019，233(9)：2323.

[12] ROTHER C，KOLMOGOROV V，BLAKE A. GrabCut：interactive foreground extraction using iterated graph cuts [J]. ACM Transactions on Graphics，2004，23(3)：309.

[13] LECUN Y，BOTTOU L，BENGIO Y，et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE，1998，86(11)：2278.

[14] SIMONYAN K，ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2019-09-08]. https://arxiv.org/pdf/1409.1556.pdf.

[15] SZEGEDY C，VANHOUCKE V，IOFFE S，et al. Rethinking the inception architecture for computer vision[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Las Vegas：IEEE，2016：2818.

[16] CHOLLET F. Xception：deep learning with depthwise separable convolutions[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Honolulu：IEEE，2017：1800.

[17] SANDLER M，HOWARD A，ZHU M，et al. MobileNetV2：inverted residuals and linear bottlenecks[C]//2018 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Salt Lake City：IEEE，2018：4510.