

优化卷积网络及低分辨率热成像的夜间人车检测与识别

于龙姣, 于博, 李春庚, 安居白

(大连海事大学 信息科学技术学院, 辽宁 大连 116026)

摘要: 夜间环境下人车的检测与识别在自动驾驶, 安防等领域具有重要意义。本文提出使用性价比较高的低分辨率红外热成像摄像机拍摄的图像来进行夜间的人车检测与识别, 并根据图像独特的性质对 Faster RCNN 网络进行了优化。增加多通道卷积层来适应热成像图像的灰度特性。使用全局平均池化层来适应较少的图像及类别数量, 增加批标准化层来防止加深加宽网络后可能出现的梯度消失或爆炸。使用在城市夜间环境中采集的 2000 张低分辨率热成像图像对网络进行训练与测试, 平均准确识别率达到 71.3%。相比于传统的检测手段, 本组合方法在真实的场景中取得了较好的识别效果, 同时提升了准确识别率, 有效解决了夜间环境下人车检测与识别的问题, 鲁棒性及应用价值较强。

关键词: 自动驾驶; 夜间环境; 人车检测与识别; 红外热成像; Faster RCNN

中图分类号: TP183 文献标识码: A 文章编号: 1001-8891(2020)07-0651-09

Detection and Recognition of Persons and Vehicles in Low-Resolution Nighttime Thermal Images Based on Optimized Convolutional Neural Network

YU Longjiao, YU Bo, LI Chungeng, AN Jubai

(School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China)

Abstract: The detection and recognition of persons and vehicles in the nighttime environment is highly important in the fields of self-driving cars and security. This paper proposes to use images taken by a cost-effective low-resolution infrared thermal imaging camera. We optimize the faster region-based convolutional neural network according to the unique nature of the images. A multi-channel convolution layer is added to accommodate the grayscale characteristics of thermographic images. We use a global average pooling layer so that fewer images and categories are needed, and we add batch normalization layers to prevent the appearance of exploding or vanishing gradients after the network is widened. The network is trained and tested using 2000 low-resolution thermal images collected in an urban nighttime environment. The average accurate recognition rate is 71.3%, indicating that the method effectively solves the problem of detection and recognition of persons and vehicles in the nighttime environment. The stickiness value and application potential are high.

Key words: self-driving car, night time environment, detection and recognition of persons & vehicles, infrared thermal imaging, Faster RCNN

0 引言

夜间环境下的人车检测与识别一直是计算机视觉领域中一项非常重要的研究工作。2018年Uber无人车发生撞人事故, 当地警察局长 Sylvia Moir 透露: “在观看过车载录像之后, 我们发现无论处于哪种模式(自动驾驶模式或人工驾驶模式), 本次碰撞都难以避免, 因为受害人是从暗处突然闯入机动车道的。”

调查报告显示在撞击发生的前 6s, 激光雷达的决策过程发生了误判, 而可见光摄像机由于处在黑暗环境中, 无法检测到行人, 也没有发挥任何警示作用^[1]。在现有的夜间安防监控中, 大部分的红外摄像机, 受光照条件和照射距离的影响, 极易产生噪声及过度曝光的问题, 导致不能及时发现可疑人员和车辆。因此, 在夜间环境中寻找一种有效检测与识别人车的途径显得尤为重要。本文所使用的是红外热成像摄像机拍

收稿日期: 2019-11-19; 修订日期: 2020-04-24.

作者简介: 于龙姣(1995-), 女, 硕士研究生, 主要从事计算机视觉, 利用深度学习的多目标检测算法研究。E-mail: yulongjiao@dmlu.edu.cn

基金项目: 国家自然科学基金面上项目(61471079)。

摄的图像^[2]，其不同于红外摄像机。红外摄像机使用不加装红外线过滤片的镜头，并利用红外 LED 点阵发射出的近红外光源照射来呈现出图像。红外热成像摄像机又称热像仪，其原理为通过镜头镜片材质选择过滤掉绝大多数的光线，只允许较窄取值范围的远红外自发光照射到摄像机传感器从而达到成像效果^[3]。热成像摄像机不受外界光照条件影响，只取决于物体本身的热量大小，因此可在夜间环境下拍摄到人体、车辆等自身可以散发出热量的目标，不会像红外摄像机那样将很多细节呈现出来，一定程度上减小了初始图像的噪声，热成像摄像机还具有探视距离较远的优点。上述 3 种摄像机的参数已在表 1 列出。经过综合考虑，采用热成像摄像机进行夜间人车的检测与识别在自动驾驶、安防等领域中具有良好的应用前景。

图 1 的 3 幅图像是用不同相机对同一街景的拍摄效果。在(a)图像中，使用可见光摄像机拍摄，即使有路灯照射，可我们几乎看不到远处有行人出现。在(b)图像中，使用红外摄像机拍摄，画面大致可以看出存在行人与车辆，但是受路灯等其他光照影响，产生了光斑噪声，这会影响系统的判断。在(c)图像中使用热成像摄像机拍摄，可以观察到近处的车辆和较远处的行人，因不受环境光照等影响，图像噪声较少。

1 检测与深度学习

近些年来，一些专家学者们也对夜间黑暗环境下物体的检测与识别进行了研究。Urban Meis 使用基于统计分类器的像素点、区域的分割算法和多项式分类器来检测和分类热成像图像中的对象。第一个分类器找到有潜在对象的感兴趣区域，基于区域的分割算法

用于重新分割这些 ROI (Region Of Interest)，二次多项式分类器确定对象的类型，重新分类模块进行最终检测正确与分类错误的改进^[4]。Yunyun Cao 提出了一种改进的局部二值模式 (Local Binary Pattern) 特征提取方法，用于夜间黑暗环境下的行人检测，方法是：
①利用幅度分量对 LBP 码进行加权；②使用多分辨率降低噪声的影响；③利用多尺度信息来获得灰度模式的更多共现信息。该方法可以克服部分夜间黑暗环境中低对比度、图像模糊和图像噪声的问题^[5]。Thou-Ho (Chao-Ho) Chen 利用颜色变化和前灯信息的特征来实现夜间交通场景中的车辆分割。从初始物体掩模中尽可能地减少地面的照明来获得较好的结果。使用前灯信息实现车辆流量的统计，而不是使用整个车身。实验结果表明，在中等车流量的条件下，驾驶员通常会在黑暗环境中打开大灯，此时便可以检测到车辆^[6]。

近几年随着人工智能的火热，深度学习越来越多地用于计算机视觉领域，深度学习通过组合低层特征形成更加抽象的高层来表示属性类别或特征，以发现数据的分布式特征表示，继而学习样本数据的内在规律和表示层次。因此采用深度卷积神经网络来进行图像的检测与识别可以取得非常好的效果^[7]。经过 RCNN (Regions with CNN)^[8]和 Fast RCNN^[9]的积淀，Ross B. Girshick 在 2016 年提出了新的 Faster RCNN^[10]。Faster RCNN 在结构上已经将特征抽取 (feature extraction)，建议框的生成 (proposal generate)，边框回归 (bounding box regression) 和分类 (classification) 都整合在了一个网络中，综合性能有了较大提高，在检测精度和运行速度上也优于前两种方式，因此很多目标检测识别算法都纷纷开始针对自

表 1 可见光、红外、热成像摄像机属性参数对照表

Table 1 Table of visible, infrared, thermal imaging camera property parameters comparison

	Wavelength/ μm	Spectral properties	Night vision distance/m	Night vision performance
Visible light camera	0.4-0.75	Visible light	50 \pm	Worse
Infrared camera	0.75-1.5	Infrared	100 \pm	Medium
Far infrared camera	8-14	Far infrared	200 \pm	Better



(a) 可见光摄像机成像效果



(b) 近红外摄像机成像效果



(c) 热成像摄像机成像效果

(a) Visible light camera imaging effect

(b) Near infrared camera imaging effect

(c) Thermal imaging camera imaging effect

图 1 不同摄像机拍摄图像对比

Fig.1 Comparison of images taken by different cameras

己的数据集对 Faster RCNN 算法进行优化改进。吴晓凤提出基于 Faster R-CNN 的手势识别算法。首先修改 Faster R-CNN 框架的关键参数,达到同时检测和识别手势的目的,然后提出扰动交叠率算法,避免训练模型的过拟合问题,进一步提高识别准确率^[11]。由于本文使用的为热成像图像,较普通可见光图像有其独特的性质与属性,我们针对这些特性在检测网络上做了更好的优化,来提高检测的精度。首先在基础的特征提取网络层后面加入了多通道的优化卷积核技术,来适应热成像图像的灰度及尺度特性。然后使用全局平均池化层代替了原有的3个全连接层,这使得网络的参数值大大减少,不仅提升了网络的计算性能,而且非常适合本文的少类别分类设置,同时有效地避免了过拟合的发生。最后,在特征提取卷积层的激活层前加入了批标准化(Batch Normalization)层,使得每个特征提取层都可以很好的控制数据的分布形态,防止反向传播时可能出现的梯度消失或爆炸,加快了网络的收敛速度。经过大量实验的验证,本文提出的算法与热成像技术的组合可有效地检测到夜间环境下的人车,在精度和速度上都有较好的表现,为计算机视觉领域夜间黑暗环境中的人车检测与识别提供了一种全新的参考方法。

2 优化的卷积网络

本文在 Faster RCNN 的基础上,针对热成像人车的检测与识别从如下3个方面做了优化,我们称之为 FIR (Far Infrared) Faster RCNN。

2.1 多通道的优化卷积核模型

卷积神经网络大多数被用于寻找图像的深度特征^[12],Faster RCNN 首先使用卷积网络提取图像的特征图,该特征图被共享用于后续 RPN(Region Proposal Network)层和 ROI Pooling 层。

在通常状况下,卷积运算是两个函数的一种数学运算,即:

$$s(t) = \int x(a)w(t-a)dt \quad (1)$$

式中: x 为输入函数; w 称为核函数; s 为输出函数; t 为当前时刻; a 为时间段中的某时刻。在涉及到图片和文本等数据时,由于数据是离散的,所以时刻 t 需要取整数,即离散形式的卷积运算为:

$$s(t) = \sum_{a=-\infty}^{+\infty} x(a)w(t-a) \quad (2)$$

对于本文的图像来说,输入的是一个二维数组 I ,核函数也是一个二维数组 K ,所以卷积公式为:

$$s(i, j) = \sum_m \sum_n I(m, n)K(i-m, j-n) \quad (3)$$

由于热成像图像在最终处理时已经去掉了颜色信息,只使用灰度值的大小来表示图像中不同的目标,所以输入图像在一定程度上损失了空间的颜色信息,且使用的低分辨率图像在目标的细节轮廓特征上也有所缺失,继而卷积过程中的 w 参数学习也随之减少。因此我们需要学习更多的尺度大小信息来提升识别的准确率^[13]。通常我们会再次加深网络来寻找更深层次的特征属性,但更深的模型意味着需要更多的参数,计算资源的消耗开始增加,模型也比较容易出现过拟合,因此盲目的增加模型的深度可能会适得其反。2014年,Google Net 提出了使用 Inception 模块^[14],它的目的是设计一种具有高性能的局部拓扑结构网络,目的是对输入图像并列的执行多个卷积运算和池化操作,最终将所有输出结果结合为某一层的特征图。其使用3个不同大小的滤波器(1×1、3×3、5×5)对输入进行卷积,此外还会执行最大池化操作。最终各个层的输出被合并起来,再传递至下一个 Inception 模块。在之后的 V2 和 V3 版本中^[15],作者为了减少特征的代表性瓶颈,又将 5×5 的卷积分解为两个 3×3 的卷积运算来提高运行速度。一个 5×5 的卷积核在消耗成本上是一个 3×3 卷积的 2.78 倍。因此这种改变在性能上会有所提升。此后又提出将 $n \times n$ 的卷积核尺寸分解为 $1 \times n$ 和 $n \times 1$ 两个卷积。例如,一个 3×3 的卷积核相当于先执行一个 1×3 的卷积核,然后再执行 3×1 的卷积核。同时还发现这种方法在成本上比使用单个 3×3 的卷积核降低了 33%。

本文为了在加深网络深度的同时可以获得更多尺度上的目标属性,使用了3种不同的卷积核来对应不同的感受野,帮助提升热成像图像的检测精度与效率,我们称作多尺度模块(Multi-Scale module, MSM)。1×1 卷积核只有一个参数,对应到特征图上就是对每一个像素点进行遍历,这样可以对特征图的细节学习的更加透彻。1×3 和 3×1 卷积核的加入使得网络不再仅是一直加深,而且加宽了网络,让网络对尺度的适应性更强。据此我们在 VGG16 网络的基础上修改了它的第四与第五卷积层,分别在这两个卷积层的3个分卷积层之后增加了 1×1, 1×3 与 3×1。经多次试验验证,对于本文的热成像图像来说,由于尺度的大小是特征提取的重要因素,若使用与 Inception 结构相同的卷积块,特征提取的效果略显不足,所以我们开创性的使用了 7×7 的卷积核,并用 1×3 和 3×1 的卷积核组合成 5×5 的卷积核大小,将这几种卷积核组合为卷积块,在提取出不同尺度的特征后,合并

输出，最终进行最大池化操作，送入下一个卷积层。相比于 VGG16^[16]加深了网络，比 VGG19 又加宽了网络，同时提升了感受野的尺度。这种优化使得网络需要学习的权重数量大幅下降，训练时间也有了一定程度的缩短。图 2 为优化后网络的结构模型。

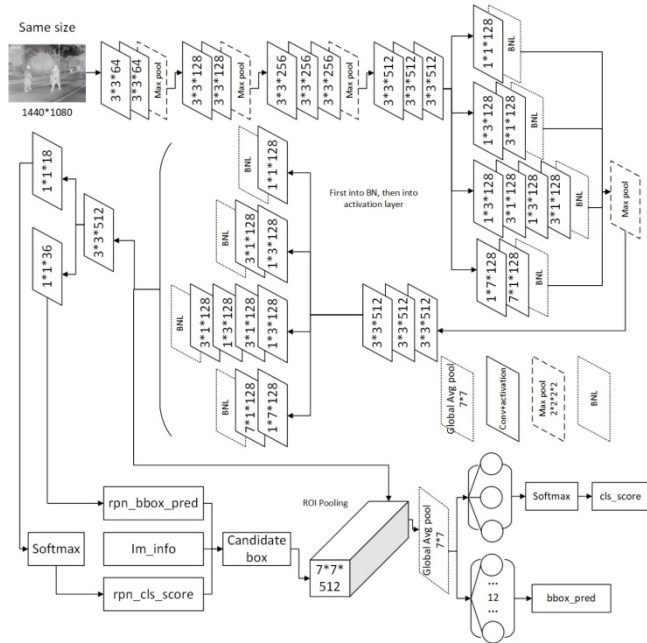


图 2 FIR Faster RCNN 示意图

Fig.2 Schematic diagram of the FIR Faster RCNN

2.2 全局平均池化层的使用

在现有的很多基于卷积神经网络的检测分类网络中，都会将最后一个卷积层得到的映射特征矢量化，然后加上全连接层来接入 Softmax 层进行逻辑回归分类。这种设计很好地将卷积层结构和传统的神经网络分类器结合起来，将卷积神经网络作为一种特征提取器，然后将得到的特征 $(x_1 \sim x_n)$ 使用式(4)(5)(6)的传统方式对其进行分类：

$$a_1 = \omega_{1,1} \times x_1 + \omega_{1,2} \times x_2 + \dots + \omega_{1,n} \times x_n + b_1 \quad (4)$$

$$a_2 = \omega_{2,1} \times x_1 + \omega_{2,2} \times x_2 + \dots + \omega_{2,4096} \times x_{4096} + b_2 \quad (5)$$

$$a_3 = \omega_{3,1} \times x_1 + \omega_{3,2} \times x_2 + \dots + \omega_{3,4096} \times x_{4096} + b_3 \quad (6)$$

式中： w 为参数权重值； b 为偏置项。

在反向传播计算时，若我们已知传递到该层的梯度 $\frac{\partial \text{loss}}{\partial a}$ ，便可以通过链式法则求得 loss 对 x 的偏导数。首先求出该层输出 a_i 对输入 x_j 的偏导数：

$$\frac{\partial a_i}{\partial x_j} = \sum_j^{4096} \omega_{ij} x_j = \omega_{ij} \quad (7)$$

再通过链式法则求得 loss 对 x_k 的偏导数：

$$\frac{\partial \text{loss}}{\partial x_k} = \sum_j^3 \frac{\partial \text{loss}}{\partial a_j} \frac{\partial a_j}{\partial x_k} = \sum_j^3 \frac{\partial \text{loss}}{\partial a_j} \times \omega_{jk} \quad (8)$$

最后对权重系数求导，由于 $\frac{\partial a_i}{\partial \omega_{ij}} = x_j$ ，所以：

$$\frac{\partial \text{loss}}{\partial \omega_{kj}} = \frac{\partial \text{loss}}{\partial a_k} \frac{\partial a_k}{\partial \omega_{kj}} = \frac{\partial \text{loss}}{\partial a_k} \times x_j \quad (9)$$

由于网络隐藏层中有许多我们无法解读的数据分布，有时设计几个全连接层可以提升卷积神经网络的分类性能，因此全连接层经常会被用在神经网络的末端，Faster RCNN 算法也不例外。但是上述运算容易发生过拟合，使得网络的泛化能力不足^[17]，且参数量过大，每层全连接都有 4096 个神经元，特别是与最后 ROI Pooling 层相连的全连接层。这大大降低了网络的运行效率。在 Network In Network 一文中^[18]，作者提出使用全局平均池化，其做法是针对每一个类别，都从特征提取层的最后一个卷积层中生成一个对应的特征图，然后对特征图上的所有点求得均值，最后将这些点直接连接到 Softmax 分类器上，代替了原来使用卷积层的特征点连接到全连接层后再连接至 Softmax 的做法。首先，这种结构使得特征图和分类器在卷积结构层面有着更强的连接响应，因此特征图可以很好地被解释成为分类置信度图。其次，由于这种做法不会使用到任何新的参数，因此不需要对参数进行优化，同时避免了过拟合的发生。此外，全局平均池化层对空间域的特征整合较好，在理解输入特征的空间特征时具有很好的鲁棒性。

对于本文设计的多通道网络模型来说，特征提取的误差主要来自两个方面：①感受野大小的变化造成的估计值方差变大；②卷积层参数误差造成估计均值的偏移。因此我们选择使用全局平均池化层 (Global-Average-Pooling, 实验中简称 GAP) 来代替全连接层，来适应我们的小样本低分辨率热成像图像。池化的结果使得最终得到的特征图被优化为一个分类置信度，使用得到的置信度神经元连接到只有 3 类 (含背景) 的 Softmax 分类器上。上述操作可以对整个网络在结构上做正则化防止过拟合，去掉了无法理解的隐藏神经元的的信息，直接赋予了每个通道实际的内在意义。此外还有效地保持了旋转、平移、伸缩的不变性，同时提高了训练速度。

2.3 批标准化层的使用

全连接层被代替后，大部分需要优化的权值参数

都集中在了前半部分的特征提取层部分。由于我们的网络在设计时进行了加深与加宽,而深层神经网络在进行非线性变换前的激活输入值 $x_1(a_1 = \omega_1 x_1 + b_1)$ 随着网络深度的加深,在训练过程中其概率分布逐渐发生偏移和端化。也就是整体分布逐渐向非线性函数取值区间的上下限两端慢慢逼近(对于 Sigmoid 激活函数来说,意味着激活输入值 x_1 会向 0 或 1 值靠近)。因此导致了反向传播时低层神经网络的梯度消失或爆炸,从而使深层神经网络训练时收敛越来越慢。而批标准化层^[19] (Batch-Normalization layer, 实验中简称 BN) 就是通过一定的规范化手段,将每层神经网络任意神经元的权重值分布强行拉回到均值为 0 方差为 1 的标准正态分布上,也就是将权重从逐渐偏离的数据分布强制拉回到比较标准的分布,这样使得激活输入值可以落在非线性函数对输入比较敏感的区域。此时参数上较小的变化,便能在损失函数上体现出较大的变化。从而使梯度变大,避免梯度消失或爆炸的问题产生,而且梯度变大意味着学习收敛速度快,能有效地提升训练速度。具体过程如下:

先对小批量送入网络训练的 d 维参数 $x=(x(1)\cdots x(d))$ 进行单独标准化,使其具有零均值和单位方差。

$$\hat{x}^{(k)} = \frac{x^{(k)} - E[x^{(k)}]}{\sqrt{\text{Var}[x^{(k)}]}} \quad (10)$$

式中: E 为求取其均值; Var 为求取其方差。

然后要确保插入到网络中的变换可以表示恒等变换。因此对于每一个激活 $X^{(k)}$, 都会引入成对的参数 $\gamma^{(k)}$ 和 $\beta^{(k)}$, 它们会归一化和标准化输入值。即:

$$y^{(k)} = \gamma^{(k)} x^{(k)} + \beta^{(k)} \quad (11)$$

因此首先根据式(12)求出输入值的均值,然后根据式(13)求出输入值的方差,根据式(14)将输入值标准化后,训练参数 $\gamma^{(k)}$ 和 $\beta^{(k)}$ 的值,最终使其成为一个批标准化的恒等映射。

$$\mu_B \leftarrow \frac{1}{m} \sum_{i=1}^m x_i \quad (12)$$

$$\sigma_B^2 \leftarrow \frac{1}{m} \sum_{i=1}^m (x_i - \mu_B)^2 \quad (13)$$

$$\hat{x}_i \leftarrow \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} \quad (14)$$

$$y_i \leftarrow \gamma \hat{x}_i + \beta \equiv \text{BN}_{\gamma, \beta}(x_i) \quad (15)$$

式中: x_i 为输入参数; μ_B 为其均值; σ_B^2 为其方差; ε 为偏置项; y_i 为最终输出; BN 为批标准化操作。

在将批标准化层加入我们设计的网络实验时发

现,根据式(16)~(21)的链式法则进行反向传播计算损失值时,若将其加在特征网络基本层的所有卷积层后,网络虽能有效地快速收敛,但在测试的准确率上却没有突出的表现。经研究发现,由于原有基本特征提取层的训练值使用的为预训练模型的参数值,在数据初始分布上已经有了比较好的标准化,再次进行本操作意义不大,且可能对数据分布产生噪声影响。因此我们修改为只在优化后加深与加宽的多通道网络上使用批标准化功能。此时,训练后的梯度分布便可以较好地反映到需要调整参数较多的多通道卷积层中,同时也不会影响到原有预训练网络模型的参数分布。

$$\frac{\partial l}{\partial \hat{x}_i} = \frac{\partial l}{\partial y_i} \cdot \gamma \quad (16)$$

$$\frac{\partial l}{\partial \sigma_B^2} = \sum_{i=1}^m \frac{\partial l}{\partial \hat{x}_i} (x_i - \mu_B) \cdot -\frac{1}{2} (\sigma_B^2 + \varepsilon)^{-\frac{3}{2}} \quad (17)$$

$$\frac{\partial l}{\partial \mu_B} = \sum_{i=1}^m \frac{\partial l}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma_B^2 + \varepsilon}} \quad (18)$$

$$\frac{\partial l}{\partial x_i} = \frac{\partial l}{\partial \hat{x}_i} \cdot \frac{1}{\sqrt{\sigma_B^2 + \varepsilon}} + \frac{\partial l}{\partial \sigma_B^2} \cdot \frac{2(x_i - \mu_B)}{m} + \frac{\partial l}{\partial \mu_B} \cdot \frac{1}{m} \quad (19)$$

$$\frac{\partial l}{\partial \gamma} = \sum_{i=1}^m \frac{\partial l}{\partial y_i} \cdot \hat{x}_i \quad (20)$$

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^m \frac{\partial l}{\partial y_i} \quad (21)$$

3 实验准备与结果分析

3.1 实验图像的采集及预处理

本文使用 FLIR One Pro3 热成像摄像机进行图像的采集,选择 3 种典型的夜间场景:①明亮处,一般在城市中心或者活动广场,行人与车辆较多,光照效果好。②明暗交替处,大部分公路、街道等区域都处在这种环境,有路灯照射,但光照区域覆盖不全,此种场景行人与车辆数量适中。③黑暗处,无光源照射的街道马路,关闭的商场商店,乡村小路以及灯光昏暗的行人步道等,这些场景的行人与车辆一般较少。在这些场景中共采集了 6 段视频,每段视频半小时,使用平均时间间隔的方法,每 5 s 从视频中截图一次,抽取了 2000 张图像制作成数据集,数据集文件夹形式和公共数据集 VOC 相同^[20],使用 labelImg 工具进

行图像的标注并自动生成.xml文件,标注的类别为人与车辆(person, vehicle)。

由于本文使用的为低分辨率图像,在场景①下,可能会因为目标出现较多导致目标轮廓不清晰,这对图像标注的准确度造成影响。考虑到在有光源照射处可见光摄像机还会捕捉到一些图像信息,本文提出了运用可见光摄像机拍摄到的图像进行边缘检测,然后将热成像图像和进行边缘检测后的图像进行融合,在融合图像上做标记,最后将位置与分类信息存入文件,有效解决了标记困难的问题,从而为低分辨率图像的训练任务提供了先行条件。下面详细介绍一下融合方法中涉及到的图像处理过程。

3.1.1 可见光图像的边缘检测

在处理过程中,若使用可见光图像直接进行融合,融合后的图像会有更多的噪声导致无法标记。所以本文提出对可见光图像进行边缘检测,然后融合到热成像图像上的方法,有效地在热成像图像上呈现出了清晰的目标轮廓。图3列出了常用的3种边缘检测算法在本文图像上的效果。通过对比,Sobel对人物及车辆产生较好的边缘检测效果,同时,由于其引入了局部平均,使其受噪声的影响也较小,效果好。Laplace对噪声具有无法接受的敏感性,检测效果不好。Canny是目前理论上相对最完善的一种边缘检测算法,但在检测人物与车辆细节上有一些缺失,效果较好,但细节不如Sobel。综上所述,我们最后选择Sobel边缘检测算法来进行图像融合。

3.1.2 图像融合标记

首先找到肉眼无法分辨轮廓或类别的热成像图像,根据名称对应找到可见光摄像机拍摄的图像。使用可见光图像进行Sobel边缘检测得到边缘检测图,因为两个摄像机在同时拍摄时的物理位置上有一定距离,所以在融合前需要找到一个合适的偏移量,然后根据此参数对图像进行位置偏移。由于热成像摄像机和可见光摄像机的物理距离是固定的,所以找到此参数后便可反复使用。将偏移好的边缘检测图像与成像不清晰的热成像图像进行叠加融合,便可得到较清晰的融合图像。在融合时,使用边缘检测图像进行左右移动来匹配热成像图像,成功匹配后进行标记,由于热成像图像和融合图像的大小与人车的相对位置都已对应,所以可将分类与位置信息直接存入.xml文件。

在图4(a)中,购物广场的环境光照较充足,可见光摄像机能捕捉到一些图像信息,可用来辅助热成像图像的标注工作。(b)中因行人较多,在热成像图像中会有重叠和模糊现象的存在,导致看不清到底有几个人。(c)中因为两个摄像机在安装时会有物理上的距离,将边缘检测后的可见光图像和热成像图像融合后,出现了位置不对应的情况,比如图中圈出人的轮廓与实际位置不对应。(d)中经过偏移融合后,我们可以清晰地看出图像中每个目标的轮廓,圈中人的位置也可以正确对应,大大提高了图像标注的准确度与效率。



(a) 可见光图像 (b) Sobel 边缘检测图像 (c) Laplace 边缘检测图像 (d) CANNY 边缘检测图像
(a) Visible light image (b) Sobel edge detection image (c) Laplace edge detection image (d) CANNY edge detection image

图3 常用边缘检测算法对比 Fig.3 Comparison of common edge detection algorithms



(a) 可见光图像 (b) 热成像图像 (c) 未偏移的融合图像 (d) 偏移操作后的融合图像
(a) Visible light image (b) Thermal image (c) Unshifted fused image (d) Offset fused image

图4 不同形式的图像对比 Fig.4 Comparison of different forms of images

3.2 测试环境及训练参数设置

本文使用一块 Nvidia GTX1080Ti 11G 显存的 GPU 进行实验,实验环境为 Ubuntu16.04+CuDa8.0+Cudnn5.1+TensorFlow1.2.0。数据集共有 2000 张图像,采用平均随机分布的方法从中抽取 1400 张图像作为训练集,从剩下的 600 张中用相同的方法抽取 200 张图像作为测试集,剩余 400 张为验证集。经多次实验,在训练 40000 次后 loss 值基本稳定收敛,故将训练的 次数设置为 40000,学习率开始设置为 0.005,随后每 10000 次衰减 50%。图像大小固定尺寸至 1440 × 1080。BN 层的 decay 参数设置为 0.9,将基础特征

网络的前两层 is training 设置为 False。

3.3 结果分析

首先验证热成像方法的有效性,在测试时,对于可见光图像使用 Faster RCNN 网络 (VGG16) 进行测试,对于热成像图像,使用 FIR Faster RCNN 网络测试,测试图像均为同一场景下使用不同摄像机拍摄的,且未被训练。根据图 5 的检测效果来看,本文提出的网络在热成像图像中可有效地检测与识别人车目标,且在类似场景(3)的环境下,检测效果显著优于可见光图像。









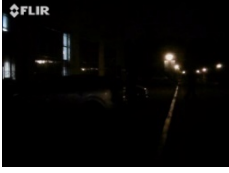



	Faster RCNN (VGG16、Visible light image detection)	FIR Faster RCNN (Small sample low-resolution thermal imaging image)	Faster RCNN (VGG16、Small sample low-resolution thermal imaging image)	Faster RCNN (VGG19、Small sample low-resolution thermal imaging image)
Scenes ①				
Scenes ②				
Scenes ③				

图 5 检测效果对比图 Fig.5 Comparison of inspection results

其次验证优化后网络的可靠性,图 6 所示的依据 Tensor Board 统计数值画出的曲线我们可以得到,在训练的过程中,本文提出的优化网络最终的总体损失终值为 0.11,各个参数的损失值都可以随着训练次数的增加而逐渐收敛到一个稳定的数值。表 2 的数据说明了本文引入与设计各个模块对于模型最终结果的影响程度。多通道卷积核有效提升了模型的预测精度;对于本文的小样本数据集,全局平均池化可显著优化模型的过拟合能力;批标准化的使用使得模型可以快速收敛并得到模型的最优结果。图 5 中,本文方法可较好地检测出目标,边框回归位置也比较精准,特别是在少样本的车类别检测与识别中,相对其他两种网络表现较好。

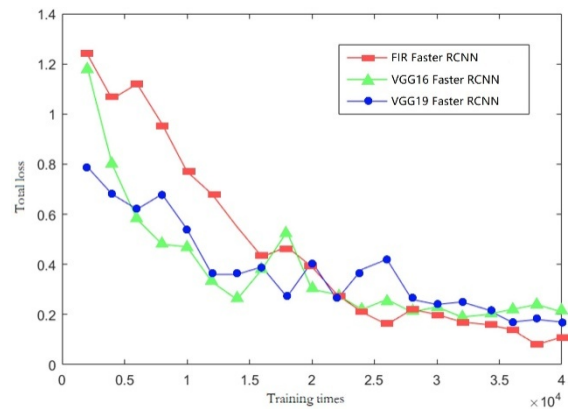


图 6 网络训练总体收敛曲线对比 Fig.6 Network training overall convergence curves comparison

表 2 各模块性能对比

Table 2 Performance comparison of each module

Model	Train time/min	Test set		mAP
		Person(AP)	Car(AP)	
Baseline(VGG16)	158.5	68.8	67.3	68.05
Baseline+1*1(256)+3*3(256)	155.7	67.6	66.4	67
Baseline+MSM	156.9	72.2	68.1	70.15
Baseline+MSM+GAP	151.3	73.5	68.5	71
Baseline+MSM+GAP+BNL	145.4	73.9	68.7	71.3

3.4 评价指标

使用平均精确度 (Average Precision) 指标来对所有测试集图像进行分析。针对数据集 D 和学习器 f 而言:

1) 错误率: 分类错误的样本数占总样本的比例, 即:

$$E(f; D) = \frac{1}{m} \sum_{i=1}^m I(f(x_i) \neq y_i) \quad (22)$$

2) 精度: 分类正确的样本数占总样本的比例, 即:

$$ACC(f; D) = \frac{1}{m} \sum_{i=1}^m I(f(x_i) = y_i) = 1 - E(f; D) \quad (23)$$

对于本文的检测类别 (人) 来说, 在测试集中的一张图像里, 精确度(Precision)=此图像识别正确的人的数量/此图像标签中人的总数。平均精确度=对含有人的图像精确度求和/含有人的图像总数。总体平均精确度(mean Average Precision)=对人和车的平均精确度求和/2。最后在自制的数据集上使用两种方法分别进行测试。

表 2 和表 3 的数值可以量化分析识别的准确率, 验证网络的识别效果。根据表中数据我们可以看出本文方法在平均准确度上高于 VGG16 及 VGG19。VGG19 虽然在人的识别准确率上较高, 但其受小样本目标分布不均衡的影响较大, 在车的分类准确率上表现不佳, 不具有泛化能力。最终上述各项指标的结果证明了本文网络设计方案的可行性及泛化能力。

从图 5 和表 3、4 中可以看出, 本文设计的优化网络较先前方法可较好地检测出目标行人, 但由于数据图像的分辨率较低, 部分与人体温度接近的背景目标与人体边界处并不能有效地在图像中呈现, 导致目标回归框的定位仍有偏差。场景①中由于行人目标较小, 存在漏检的情况。针对上述问题, 在下一步的研究工作中, 考虑设计一种基于深度学习的显著图融合模型来增强远红外图像中的行人目标, 并尝试使用超分辨率网络来对低分辨率的热成像数据进行分辨率增强, 使其尽可能的被锐化, 提升对于小目标行人的

检测识别率。同时在我们研究的过程中, 其他机构也发布了一些分辨率较高的热成像图像, 我们会在此基础上继续深入的研究。

表 3 类别 AP 值

Table 3 Class AP Values

	FIR Faster RCNN	Faster RCNN	
		VGG16	VGG19
Person	73.9	68.8	74.8
Car	68.7	67.3	66.8

表 4 mAP 值及效率

Table 4 mAP values and detection times

	FIR Faster	Faster RCNN	
	RCNN	VGG16	VGG19
Mean average precision(mAP)	71.3	68.05	70.8
Train time(min)	145.4	158.5	170.8
Fps(f/s)	0.2	0.3	0.45

4 结论与展望

针对传统方法在夜间环境下人车检测与识别效果不佳的情况, 本文提出了使用小样本低分辨率热成像图像和优化卷积网络组合的方式来提升检测与识别的精度。分别在明亮, 明暗交替和黑暗 3 种典型的夜间场景进行了实验。根据实验结果显示, 优化后的网络可以较好地检测到物体并准确分类, 实际效果明显优于可见光图像。在之后的工作中, 我们也会继续寻找优化方法来提升识别准确率, 助力热成像技术在计算机视觉领域里的普及。综上所述, 使用小样本低分辨率红外热成像图像和优化卷积神经网络来进行夜间环境下的人车检测与识别取得了良好的效果, 在自动驾驶和安防等领域具有较高的普适性和实用价值。

参考文献:

[1] Holstein T, Dodig-Crnkovic G, Pelliccione P. Ethical and Social Aspects of Self-Driving Cars[EB/OL]. *Arxiv preprint*, 2018, arXiv:1802.04103.

- [2] Gaussorgues G, Chomet S. *Infrared Thermography*[M]. Springer Science & Business Media,2012, 5:379-395
- [3] Beard J. Introduction to infrared thermography[J]. *Lecture Presented at IT*, 2007, **570**: 7-14.
- [4] Meis U, Ritter W, Neumann H. Detection and classification of obstacles in night vision traffic scenes based on infrared imagery[C]//*Proceedings of the 2003 IEEE International Conference on Intelligent Transportation Systems*, 2003, 2: 1140-1144.
- [5] Cao Y, Pranata S, Nishimura H. Local binary pattern features for pedestrian detection at night/dark environment[C]//2011 18th *IEEE International Conference on Image Processing. IEEE*, 2011: 2053-2056.
- [6] Chen T H, Chen J L, Chen C H, et al. Vehicle detection and counting by using headlight information in the dark environment[C]//*Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2007). IEEE*, 2007, **2**: 519-522.
- [7] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks[C]//*Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014: 1725-1732.
- [8] Lenc K, Vedaldi A. R-cnn minus r[EB/OL]. *ArXiv preprint*, 2015, arXiv:1506.06981.
- [9] Girshick R. Fast r-cnn[C]//*Proceedings of the IEEE international conference on computer vision*, 2015: 1440-1448.
- [10] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2017(6): 1137-1149.
- [11] 吴晓凤, 张江鑫, 徐欣晨. 基于 Faster R-CNN 的手势识别算法[J]. *计算机辅助设计与图形学学报*, 2018, **30**(3): 468-476.
- WU Xiaofeng, ZHANG Jiangxin, XU Xinchun. Hand Gesture Recognition Algorithm Based on Faster R-CNN[J]. *Journal of Computer-Aided Design & Computer Graphics*, 2018, **30**(3): 468-476.
- [12] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//*Advances in Neural Information Processing Systems*, 2012: 1097-1105.
- [13] Cireřan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification[EB/OL]. *Arxiv Preprint*, 2012, arXiv:1202.2745.
- [14] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//*Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition*, 2015: 1-9.
- [15] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//*Proceedings of the IEEE Conference On Computer Vision And Pattern Recognition*, 2016: 2818-2826.
- [16] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. *Arxiv Preprint*, 2014, arXiv:1409.1556.
- [17] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//*Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015: 3431-3440.
- [18] LIN M, CHEN Q, YAN S. Network in network[EB/OL]. *ArXiv preprint*, 2013, arXiv: 1312.4400.
- [19] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[EB/OL]. *ArXiv preprint*, 2015, arXiv: 1502.03167.
- [20] Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (voc) challenge[J]. *International Journal of Computer Vision*, 2010, **88**(2): 303-338.