基于红外与雷达的夜间无人车场景深度估计

姚广顺[1,2]，孙韶媛[1,2]，方建安[1,2]，赵海涛[3]

[1]东华大学信息科学与技术学院，上海 201620；

[2]东华大学数字化纺织服装技术教育部工程研究中心，上海 201620；

[3]华东理工大学信息科学与工程学院，上海 200237

摘要 单目红外图像的深度估计是夜间无人车场景理解的关键，针对夜间无人车场景的深度估计，提出一种基于深度卷积-反卷积神经网络的深度估计方法。将红外图像和雷达距离数据作为深度卷积-反卷积神经网络的输入，并将深度估计问题转化为像素级分类任务进行深度估计模型的训练。将雷达的距离数据根据深度值的范围量化为与红外图像像素一一对应的离散值并对其做标记，然后训练过程采用分类的思想解决深度估计问题。实验结果表明，利用训练得到的深度估计模型对夜间无人车获取的红外图像进行深度估计的时间为 0.04 s/frame，达到了实际应用中的实时性要求。

关键词 图像处理；红外图像；深度估计；卷积神经网络；反卷积

中图分类号 TN219 **文献标识码** A

doi：10.3788/LOP54.121003

Depth Estimation of Night Driverless Vehicle Scene Based on Infrared and Radar

Yao Guangshun[1,2], Sun Shaoyuan[1,2], Fang Jian′an[1,2], Zhao Haitao[3]

[1] *College of Information Science and Technology, Donghua University, Shanghai* 201620, *China*;

[2] *Engineering Research Center of Digitized Textile & Fashion Technology, Ministry of Education, Donghua University, Shanghai* 201620, *China*;

[3] *School of Information Science and Engineering, East China University of Science and Technology, Shanghai* 200237, *China*

Abstract Depth estimation of monocular infrared image is a key to scene understanding of night driverless vehicle. Aiming at the depth estimation of night driverless vehicle scene, a depth estimation method based on the deep convolution-deconvolution neural network is proposed. Infrared images and radar depth data are fed to the deep convolution-deconvolution neural network. The depth estimation problem is transformed to a pixel-wise classification task in the training of the depth estimation model. The radar depth values are quantized into discrete bins corresponding to the pixels of infrared image and the bins are labeled according to their depth range. The deep convolution-deconvolution neural network based depth estimation model is trained by classifying each pixel to the corresponding depth. The experimental results show that the depth estimation time is 0.04 s/frame, which use the depth estimation model to estimate the scene depth information of infrared image captured by the night driverless vehicle, and the real-time requirement in practical applications is reached.

Key words image processing; infrared image; depth estimation; convolutional neural networks; deconvolution

OCIS codes 100.4996; 100.5010; 040.3060

收稿日期：2017-06-01；**收到修改稿日期**：2017-07-05

基金项目：国家自然科学基金(61375007)、上海市科委基础研究项目(15JC1400600)

作者简介：姚广顺(1992—)，男，硕士研究生，主要从事红外图像处理与模式识别方面的研究。E-mail：yaoguangshun_64@163.com

导师简介：孙韶媛(1974—)，女，博士，教授，博士生导师，主要从事夜视机器视觉方面的研究。E-mail：shysun@dhu.edu.cn(通信联系人)

1 引 言

从一幅和多幅图像中恢复场景的深度信息是机器视觉领域的基本研究课题[1],在机器人运动控制、场景理解、场景重建等方面具有重要的应用。

针对彩色图像,深度估计技术主要利用双目深度信息和图像连续序列的三角测量。而传统的单目深度估计技术利用几何假设和手动选取的特征比如尺度不变特征变换(SIFT)、分层梯度方向直方图(PHOG)、线性滤波器(GIST)、纹理等。较经典的是"由阴影恢复形状"[2]。Saxena 等[3]采用马尔可夫场和线性回归学习图像集的特征实现对图像的深度估计,取得了较好的效果。

深度学习被认为是一项机器学习领域的突破性技术,通过发现多层次的特征表达得到高级的特征来抽象表达数据。目前也有一些方法将深度学习用在了图像的深度估计上面。这一领域的突出工作是由 Eigen 等[4]完成,他们采用多尺度的深度卷积网络实现了图像的深度估计。在此之后,Liu 等[5]引入条件随机场(CRF),经过深度卷积网络学习,不需要任何先验知识和信息完成了通用场景的深度估计。Mclaughlin 等[6]引入时间递归神经网络和深度卷积网络相结合的方式实现了不同场景的深度估计。

上述前人的研究工作都是基于回归思想且在彩色图像范围内,而夜间无人车所获取的单目红外图像完全不同于可见光图像,其无色彩且纹理简单[7],对比度和信噪比低[8]且图像模糊[9],目前国内外对于夜视图像的深度估计研究还不够深入,尚无较好的解决方法。席林等[10]采用支持向量机模型学习设定的图像特征。该模型能从整体上估计单目红外图像的深度信息,该方法的缺点是结果不太准确,计算慢。沈振一等[11]采用马尔可夫场模型学习超像素面板参数和深度信息的关系,从而实现估计给定超像素块的深度信息,该方法的缺点是需要人工选取特征,无法挖掘图像深层特征信息。

以上方法通过训练模型学习手动选取的红外图像特征,但无法提取深层次特征以及实时估计场景深度信息。为了改进这种情况,本文提出了利用卷积-反卷积网络的基于红外图像和雷达数据的夜间无人车场景深度估计方法。该方法通过深度卷积-反卷积网络对红外图像和雷达数据训练得出深度估计模型,其中卷积神经网络主要用于特征提取,反卷积网络主要用于雷达数据的匹配。实验结果证明,该方法的准确率高并且能够实时地估计夜间无人车场景深度。

2 卷积-反卷积神经网络

2.1 特征表示

红外图像反映的是场景的温度分布,但是场景中物体的热辐射量可能相似,造成红外图像模糊,局部细节不足[12],为了能更准确描述图像特征,采取卷积神经网络来提取图像的特征,该方法的优点是:1) 不需要任何人工设计的特征,直接从原始图像上提取特征;2) 不对场景做任何假设,即不需要引入语义信息,也不需要约束场景的结构。

一个典型的卷积神经网络包含卷积层、激励层和池化层多个阶段,每个阶段的输入和输出的数组集合叫特征图。网络最后输出的特征图可以看作是从输入图像的所有位置上提取出来的特定特征。前几个阶段是由卷积层和池化层组成的降采样过程,卷积层的单元被组织在特征图中,通过一组滤波器连接到上一层的特征图中的一个局部块,然后这个局部的加权和被传递给一个非线性激活函数。在这里特征图的神经元使用相同的过滤器,不同层的特征图使用不同的过滤器,对于图像来说,这样做有两个好处:1) 像素点邻域内的特征都是相关的,可以形成非常重要的局部特征;2) 不同位置局部特征具有差异性,因此不同位置的神经元能够共享权值。

选取去掉全连接层的 VGG16[13]作为卷积神经网络,如图 1 所示,该网络可以直接从训练样本中学习特征,避免了手动选取特征。

对于深度估计的任务来说,给定一幅图像 X,目的是估计每个像素点的深度,所以深度估计对像素的位置精确度要求很高。为了解决这个问题,文献[14]受到深度学习层次可视化的启发,将多阶段的输出进行合并来强化结果,在合并的过程中采用简单的双线性插值的方法完成反卷积操作。周围 4 个点的像素值大小通过插值的方法得到中间点的像素值,填充原来被池化的层,从而得到原图大小[13]。

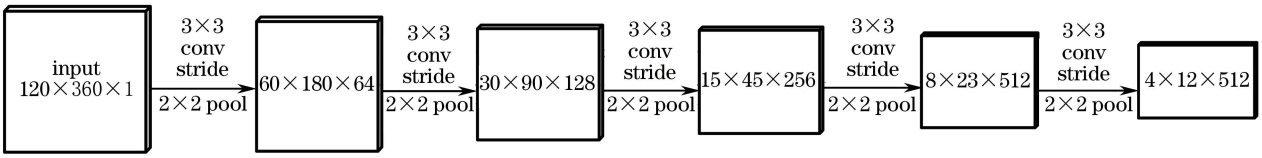

图 1 卷积神经网络结构图

Fig. 1 Structure of convolutional neural network

假设图像 X 作为卷积神经网络的输入，卷积神经网络 f 包含 L 个阶段，可训练的卷积核参数为 $\theta_f=(\boldsymbol{W},\boldsymbol{b})$，则输入 X 经过卷积神经网络 f 后的输出特征向量可表示为

$$f(X,\theta_f)=\mathrm{nonl}(\boldsymbol{W}_l\boldsymbol{H}_{l-1}+\boldsymbol{b}_{l-1}), \tag{1}$$

式中 nonl(•)表示逐点的非线性激活函数，$\boldsymbol{H}_{l-1}$ 表示第 $l-1$ 阶段的输出；$\boldsymbol{W}_l$ 表示第 l 和第 $l-1$ 阶段的连接矩阵，由卷积核的参数组成；$\boldsymbol{b}_l$ 为偏置参数向量。将第 l 个阶段的输出表示为

$$\boldsymbol{H}_l=\mathrm{pool}[\mathrm{nonl}(\boldsymbol{W}_l\boldsymbol{H}_{l-1}+\boldsymbol{b}_l)],\ l\in\{1,\cdots,L-1\}, \tag{2}$$

式中 $\boldsymbol{H}_l$ 为 l 层隐层单元的输出，pool(•)表示在特征图上进行池化操作。(2)式的意义是，对于上一层的输出 $\boldsymbol{H}_{l-1}$，经过卷积 $(\boldsymbol{W}_l,\boldsymbol{b}_l)$、非线性激活函数 nonl(•)和空间池化 pool(•)之后，最终得到该阶段的输出 $\boldsymbol{H}_l$。下采样层采用池化技术将小邻域内的特征点整合得到新特征，使得特征和参数减少，且池化单元具有平移不变性。

本文算法最终得到 4×12×512 的特征图，其中 4×12 表示特征图的尺寸，单位为 pixel，512 表示特征图的个数。

2.2 反卷积神经网络[15]

经过一系列卷积、池化后，深层的特征图的尺寸会远远小于原始图像的尺寸。然而雷达数据和原始图像的尺寸是同样大小的。如图 2 所示，在反卷积网络中加入反池化层，模拟池化的逆过程，在池化过程中将最大激活值的坐标记录下来，在反池化时把池化过程中最大激活值的坐标位置的值(像素)还原，而其他位置则采用补零的办法。池化时记录像素点坐标，反池化时还原像素的位置。

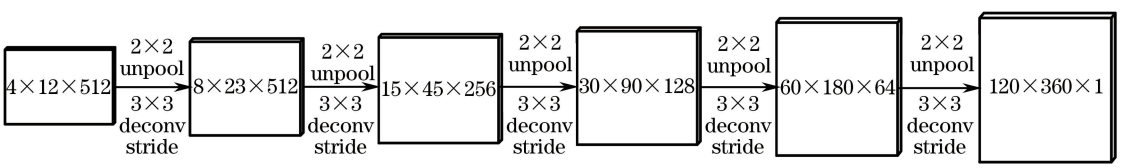

图 2 反卷积神经网络结构图

Fig. 2 Structure of deconvolution neural network

经过反池化操作的特征图扩大了两倍，但是得到的特征图是稀疏的，需要对特征图进行加密，因此本文引入了反卷积层。卷积操作将卷积核内的多个激活值连接得到一个激活值，而反卷积则是将一个激活值输出为多个激活值。同样的，反卷积操作也使用多个可学习的卷积核，并且反卷积网络和卷积网络是对称的结构，也能够获取不同层次的特征信息，较低的卷积层能够得到物体大致的形状信息，较高的卷积层则能够得到物体的细节信息。

2.3 损失函数

在网络的最后加上一个分类层就可以得到每个像素点所属各类别的概率。通过网络输出的概率图确定每个像素点所属类别，像素点所属类别为概率图中概率最大的类别。

使用多类别交叉熵损失函数来训练网络参数，确定像素点所属类别，假设类别 a 可以取 k 个不同的值，$a\in\{1,2,3,\cdots,k\}$，则：

$$\hat{c}_{ia}=\begin{bmatrix}p(a=1\mid F_i;w)\\ p(a=2\mid F_i;w)\\ \vdots\\ p(a=k\mid F_i;w)\end{bmatrix}=\frac{1}{\sum_{b\in k}\exp(w_b^{\mathrm{T}}F_i)}\begin{bmatrix}\exp(w_1^{\mathrm{T}}F_i)\\ \exp(w_2^{\mathrm{T}}F_i)\\ \vdots\\ \exp(w_k^{\mathrm{T}}F_i)\end{bmatrix}=\frac{\exp(w_a^{\mathrm{T}}F_i)}{\sum_{b\in k}\exp(w_b^{\mathrm{T}}F_i)}, \tag{3}$$

式中 w 为可训练的参数，$\hat{c}_{i,a}$ 表示预测像素点 i 属于 a 类别的概率，F_i 表示样本，w_b、$w_a \in [w_1^{\mathrm{T}}, w_2^{\mathrm{T}}, \cdots, w_k^{\mathrm{T}}]^{\mathrm{T}}$ 表示模型参数，p 表示求取概率。由(3)式可以得到预测类别与像素点实际所属类别的交叉熵损失值：

$$L = -\sum_{i \in p} \sum_{a \in k} c_{i,a} \ln(\hat{c}_{i,a}) , \tag{4}$$

式中 $c_{i,a}$ 表示实际上像素点 i 属于 a 类别的概率，k 表示类别数，p 表示所有的像素点，L 表示损失值。而若雷达数据在像素点 i 标记为 a 类，则 $c_{i,a}=1$，否则为零。

最后整个卷积网络的输出大小为 $h \times w \times k$ 的概率图，$h \times w$ 为原红外图像尺寸大小，k 为类别数量，表示每个像素点属于预定义类别的概率。

2.4 模型训练

大多数的深度学习方法采用随机梯度下降法[16]进行训练。随机梯度下降调参困难，需用适当的初始化权重，并调整学习率和动量参数。所以本文采用 L-BFGS[17]算法，L-BFGS 算法是一种拟牛顿法，是适用于大批次的并行化算法，用来最大化使用图形处理器(GPU)，能够节省大量的空间和资源。L-BFGS 算法比随机梯度下降法速度更快，收敛性更加稳定。本文使用均值为零、分布符合高斯分布 $N(0,1)$ 的权重初始化各个层权重而不必特定修改某一层的初始权重和学习率以获得较好的效果。

3 实验过程和结果分析

3.1 实验配置

本文算法使用深度学习中的 Caffe 框架[18]，且参考了该框架的一些层次结构。该算法的机器软硬件配置如下：CPU 为 Intel i5-6600，随机存取存储器(RAM)内存为 16 GB，GPU 为 Nvidia Gtx 1070，操作系统为 Ubuntu 14.04，Nvidia 并行编程和计算平台 CUDA8.0。

3.2 实验数据

采用实验室的车载红外热像仪和毫米波雷达自行采集实验数据，红外热像仪和毫米波雷达的角度略有偏差，但并不影响实验。其中红外热像仪的波段是 8～14 μm，比特数为 8 位。数据集包含训练图片 1000 张及其对应的雷达数据，测试图片 137 张，图片尺寸为 120 pixel×360 pixel。红外图像的深度值是连续的，首先将红外图像对应的深度值离散化，然后利用分类的思想去估计深度。首先将红外图像深度转换到对数空间中，然后根据类别的数量平均分割对数空间以获取对应的类别：

$$X_i = \mathrm{floor}\left(\frac{\ln d_i - \ln d_{\min}}{k}\right), \tag{5}$$

式中 X_i 表示红外图像 X 中第 i 个像素点的标签，$d_{\min}$ 表示最小的深度，d_i 表示每 i 个像素点的深度，k 表示类别数量，floor 表示向下取整函数。

将测试红外图像输入到训练好的深度卷积-反卷积神经网络模型中即可得到红外图像对应的类别，再经过对应的转换则得到算法预测深度数据：

$$d_{\mathrm{p}} = \exp(X_i \cdot k + \ln d_{\min}) 。 \tag{6}$$

3.3 实验步骤

采用所提的卷积-反卷积神经网络模型，将 1000 张红外图像及其对应的雷达标签图像作为训练集，通过不断迭代使得模型收敛，然后将整个模型参数存储下来，最后用测试图像进行测试和验证，本文算法流程如图 3 所示。在训练过程中，设置基本学习率为 0.001，迭代次数为 60000，耗时约 6 h。

3.4 实验评价指标

常见的深度估计有 4 个评价指标[4]，分别为均方根误差：

$$E_{\mathrm{RMS}} = \sqrt{\frac{1}{T}\sum_{P}(d_{\mathrm{gt}} - d_{\mathrm{p}})^2} , \tag{7}$$

平均相对误差：

$$E_{\mathrm{REL}} = \frac{1}{T}\sum_{P}\frac{|d_{\mathrm{gt}} - d_{\mathrm{p}}|}{d_{\mathrm{gt}}} , \tag{8}$$

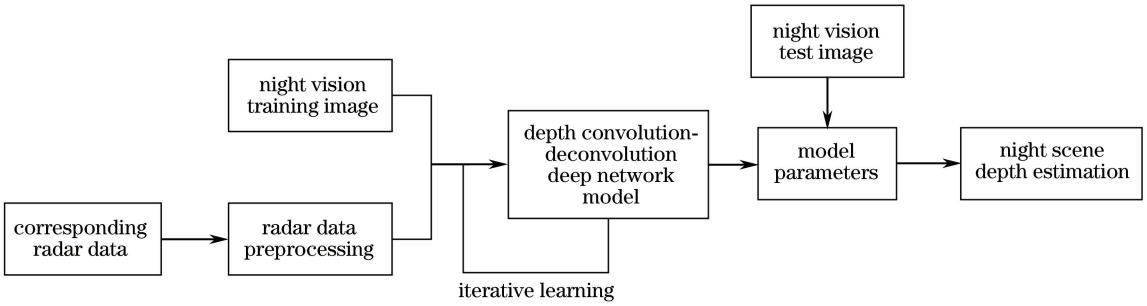

图 3 本文算法流程图

Fig. 3 Flow chart of proposed algorithm

平均 lg 误差：

$$E_{\lg}=\frac{1}{T}\sum_{P}|\lg d_{\mathrm{gt}}-\lg d_{\mathrm{p}}|, \tag{9}$$

不同阈值的准确率：

$$\delta=\max\left(\frac{d_{\mathrm{gt}}}{d_{\mathrm{p}}},\frac{d_{\mathrm{p}}}{d_{\mathrm{gt}}}\right)<1.25^{m},\ m=1,2,3, \tag{10}$$

式中 d_{gt}是像素的真实深度值，T 是所有待评价的图像的像素点总数。

3.5 实验结果及分析

将测试图像输入到训练好的模型参数中，得到深度估计图像。表 1 给出了本文算法的不同实验指标结果，表 1 中精度数据值越大效果越好，误差数据值越小效果越好，可以看出，该算法可以很好地完成红外图像的深度估计。

表 1 本文算法在不同类别时的三种评价指标结果

Table 1 Three evaluation indicators of proposed algorithm at different categories

Category	Accuracy			Error		
	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$	E_{REL}	$E_{\lg}$	E_{RMS}
10	0.682	0.873	0.952	0.330	0.126	3.215
20	0.713	0.882	0.956	0.324	0.106	3.132
50	0.725	0.895	0.958	0.320	0.097	3.093
70	0.706	0.871	0.961	0.332	0.098	3.116
100	0.692	0.862	0.961	0.336	0.102	3.119

表 2 给出了本文算法与其他经典网络模型的比较结果，可以看出，本文算法可以获取比较好的效果。

表 2 本文算法与其他经典网络模型深度估计结果的对比

Table 2 Comparison of depth estimation results between proposed algorithm and other classical network models

Algorithm	Accuracy			Error		
	$\delta<1.25$	$\delta<1.25^2$	$\delta<1.25^3$	E_{REL}	$E_{\lg}$	E_{RMS}
Proposed	0.725	0.895	0.958	0.320	0.097	3.093
Caffenet	0.624	0.852	0.891	0.342	0.129	3.932
AlexNet	0.685	0.871	0.913	0.336	0.117	3.548
FCN-VggNet	0.696	0.879	0.924	0.328	0.108	3.248

实验中对 137 幅红外图像进行测试，部分结果如图 4 所示，可以看出，本文算法的深度估计结果与图像的真实深度值比较接近。同时，本文算法用时仅需 40 ms/frame，明显少于其他方法，满足实时性。

图 5 给出了本文算法和其他经典网络模型对比的部分结果，从图中可以明显地看出，Caffenet 等网络结构只能粗略地估计出场景的总体深度，而对于场景中一些细节部分的估计则效果较差，而本文算法则能够更好地处理场景中的细节，如车辆和行人。

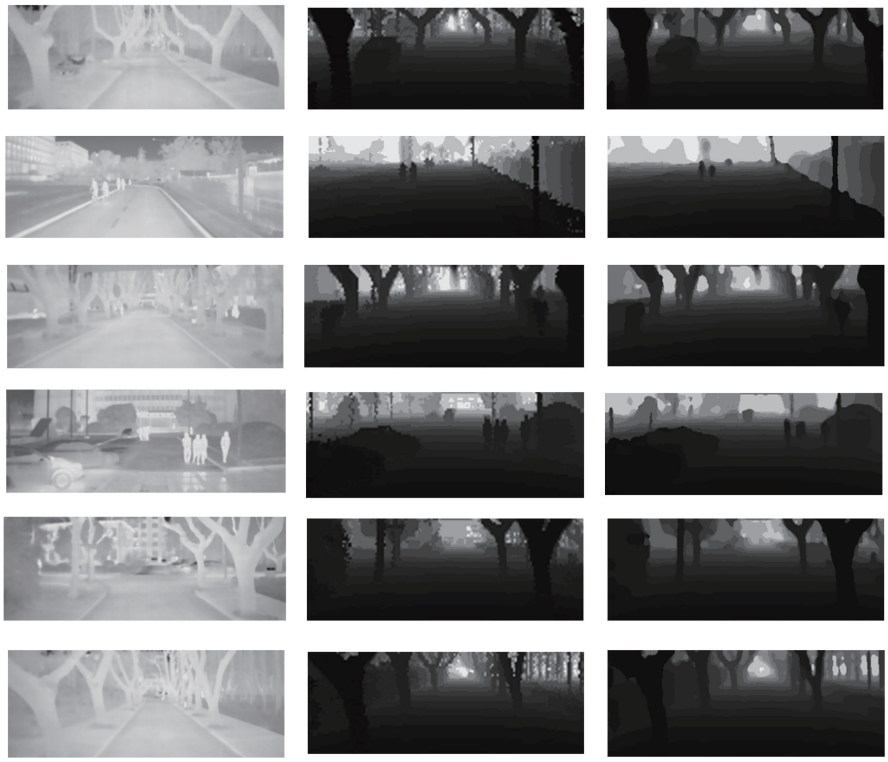

图 4 (a)红外图像;(b)真实深度图像;(c)本文算法深度估计结果

Fig. 4 (a) Infrared images; (b) real depth images; (c) depth estimation results by proposed algorithm

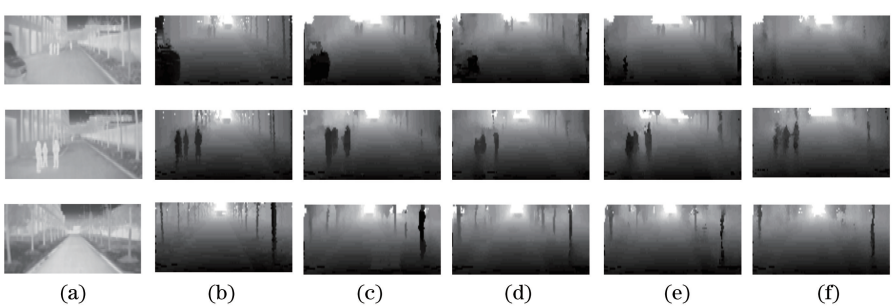

图 5 不同模型下深度估计结果。(a)红外图像;(b)真实深度图像;(c)本文算法深度估计结果;(d) FCN-VggNet 深度估计结果;(e) AlexNet 深度估计结果;(f) Caffenet 深度估计结果

Fig. 5 Depth estimation results of different models. (a) Infrared images; (b) real depth images; (c) depth estimation results of proposed algorithm; (d) depth estimation results of FCN-VggNet; (e) depth estimation results of AlexNet; (f) depth estimation results of Caffenet

4 结 论

针对夜间无人车的场景深度估计问题,提出一种基于深度卷积-反卷积网络得到深度估计模型的单目深度估计方法。利用卷积神经网络学习图像特征和深度之间的关系并得到深度估计模型。也将雷达的距离数据根据深度值的范围量化为与红外图像像素一一对应的离散值,并将其转化为像素级分类任务。实验结果表明:利用该模型能够相对准确地估计夜间无人车场景的深度信息,同时本文算法可以实时估计图像的深度。但是此模型对于红外图像深度估计的精细度有待进一步提高,后续将对红外图像深度估计做进一步的深入研究。

参 考 文 献

[1] Pentland A P. A new sense for depth of field [J]. IEEE Transactions on PAMI, 1987, 9(4): 523-531.

[2] Horn B K B. Obtaining shape from shading information[M]. New York: McGraw Hill, 1975: 115-155.

[3] Saxena A, Chung S H, Ng A Y. 3D depth reconstruction from a single still image [J]. International Journal of Computer Vision, 2008, 76(1): 53-69.

[4] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network [J]. Computer Science, 2014, arXiv: 1406.2283.

[5] Liu F, Shen C, Lin G. Deep convolutional neural fields for depth estimation from a single image[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2015: 5162-5170.

[6] Mclaughlin N, Rincon J M D, Miller P. Recurrent convolutional network for video-based person re-identification[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1325-1334.

[7] Xu Xin, Sun Shaoyuan, Sha Yujie, *et al*. A method of infrared image mosaic based on improved RANSAC[J]. Laser & Optoelectronics Progress, 2014, 51(11): 111001.
徐鑫, 孙韶媛, 沙钰杰, 等. 一种基于改进 RANSAC 的红外图像拼接方法[J]. 激光与光电子学进展, 2014, 51(11): 111001.

[8] Zou Fangyu, Sun Shaoyuan, Xi Lin, *et al*. Color stereo vision method of vehicular infrared images with depth perception[J]. Laser & Optoelectronics Progress, 2013, 50(1): 011101.
邹芳喻, 孙韶媛, 席林, 等. 具有深度视觉感的车载红外图像彩色化方法[J]. 激光与光电子学进展, 2013, 50(1): 011101.

[9] Bai Junqi, Chen Qian, Wang Xianya, *et al*. Contrast enhancement algorithm of infrared image based on noise filtering model[J]. Infrared and Laser Engineering, 2010, 39(4): 777-780.
白俊奇, 陈钱, 王娴雅, 等. 红外图像噪声滤波对比度增强算法[J]. 红外与激光工程, 2010, 39(4): 777-780.

[10] Xi Lin, Sun Shaoyuan, Li Linna, *et al*. Depth estimation from monocular infrared images based on SVM model[J]. Laser & Infrared, 2012, 42(11): 1311-1315.
席林, 孙韶媛, 李琳娜, 等. 基于 SVM 模型的单目红外图像深度估计[J]. 激光与红外, 2012, 42(11): 1311-1315.

[11] Shen Zhenyi, Sun Shaoyuan, Zhao Haitao. Three-dimensional reconstruction from monocular vehicular infrared image based on PP-MRF model[J]. Journal of Donghua University (Natural Science), 2015, 41(3): 341-347.
沈振一, 孙韶媛, 赵海涛. 基于 PP-MRF 模型的单目车载红外图像三维重建[J]. 东华大学学报(自然科学版), 2015, 41(3): 341-347.

[12] Xu Lu, Zhao Haitao, Sun Shaoyuan. Monocular infrared image depth estimation based on deep convolutional neuralnetworks[J]. Acta Optica Sinica, 2016, 36(7): 0715002.
许路, 赵海涛, 孙韶媛. 基于深层卷积神经网络的单目红外图像深度估计[J]. 光学学报, 2016, 36(7): 0715002.

[13] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2015: 3431-3440.

[14] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. Computer Science, 2014, arXiv: 1409.1556.

[15] Badrinarayanan V, Kendall A, Cipolla R, SegNet: a deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 39(12): 2481-2495.

[16] LéCun Y, Bottou L, Bengio Y, *et al*. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[17] Nocedal J, Wright S J. Numerical optimization[M]. New York: Springer, 2006.

[18] Jia Y, Shelhamer E, Donahue J, *et al*. Caffe: convolutional architecture for fast feature embedding [C]. ACM International Conference on Multimedia, 2014: 675-678.