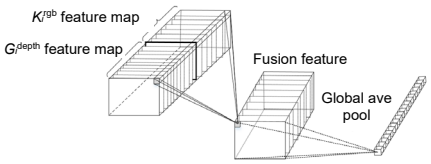
Opto-Electronic Engineering
光 电 工 程

Article

2021 年，第 48 卷，第 2 期

DOI: 10.12086/oee.2021.200069

基于改进双流卷积递归神经网络的 RGB-D 物体识别方法



李　珣[1,2]，李林鹏[1*]，Alexander Lazovik[2]，
王文杰[1]，王晓华[1]

[1]西安工程大学电子信息学院，陕西 西安 710048；
[2]格罗宁根大学伯努利实验室，格罗宁根 9747 AG，荷兰

摘要：为了提高基于图像的物体识别准确率，提出一种改进双流卷积递归神经网络的 RGB-D 物体识别算法(Re-CRNN)。将 RGB 图像与深度光学信息结合，基于残差学习对双流卷积神经网络(CNN)进行改进：增加顶层特征融合单元，在 RGB 图像和深度图像中学习联合特征，将提取的 RGB 和深度图像的高层次特征进行跨通道信息融合，继而使用 Softmax 生成概率分布。最后，使用标准数据集进行实验，结果表明，Re-CRNN 算法的 RGB-D 物体识别准确率为 94.1%，较现有基于图像的物体识别方法有显著的提升。

关键词：RGB-D 图像；结构光；物体识别；深度学习；深度图像

中图分类号：TP391.4；TP183　　**文献标志码：**A

李珣，李林鹏，Alexander Lazovik，等. 基于改进双流卷积递归神经网络的 RGB-D 物体识别方法[J]. 光电工程，2021，**48**(2): 200069

Li X, Li L P, Lazovik A, *et al*. RGB-D object recognition algorithm based on improved double stream convolution recursive neural network[J]. *Opto-Electron Eng*, 2021, **48**(2): 200069

RGB-D object recognition algorithm based on improved double stream convolution recursive neural network

Li Xun[1,2], Li Linpeng[1*], Alexander Lazovik[2], Wang Wenjie[1], Wang Xiaohua[1]

[1]School of Electronics and Information, Xi'an Polytechnic University, Xi'an, Shaanxi 710048, China;
[2]Bernoulli Institute, University of Groningen, Groningen 9747 AG, Netherlands

Abstract: An algorithm (Re-CRNN) of image processing is proposed using RGB-D object recognition, which is improved based on a double stream convolutional recursive neural network, in order to improve the accuracy of object recognition. Re-CRNN combines RGB image with depth optical information, the double stream convolutional neural network (CNN) is improved based on the idea of residual learning as follows: top-level feature fusion unit is added

收稿日期：2020-04-02；　收到修改稿日期：2020-06-13
基金项目：国家自然科学基金资助项目(61971339)；陕西省自然科学基础研究计划项目(2019JM567)；中国纺织工业联合会科技指导性项目(2018094)；大学生创新创业训练计划项目(201910709019)
作者简介：李珣(1981-)，男，博士，副教授，主要从事深度学习的多目标检测和多目标协同控制技术的研究。E-mail：leonlee527@163.com
通信作者：李林鹏(1993-)，男，硕士研究生，主要从事深度学习、计算机视觉的研究。E-mail：771613990@qq.com

into the network, the representation of federation feature is learning in RGB images and depth images and the high-level features are integrated in across channels of the extracted RGB images and depth images information, after that, the probability distribution was generated by Softmax. Finally, the experiment was carried out on the standard RGB-D data set. The experimental results show that the accuracy was 94.1% using Re-CRNN algorithm for the RGB-D object recognition, which was significantly improved compared with the existing image-based object recognition methods.

Keywords: RGB-D image; structured light; object recognition; deep learning; depth image

1 引 言

物体识别是机器视觉中的基础核心内容之一[1]。由于现实世界场景中复杂的光照和背景变化，造成现有 RGB 图像的识别算法难以满足当前智能化需求。因此，近年来结合 RGB 图像与深度图像的识别方式成为提高目标识别率的新途径。同时，对如何理解和利用两者图像的识别优势提出了新的挑战[2]。当前较多的 RGB-D 物体识别算法依靠先验知识进行目标特征的设定，已有的成果中：Lai 等人[1]定义了一种稀疏距离度量方法来快速分类；Bo 等人[3]将内核描述子扩展到深度图像，构造了较为丰富的特征；Blum 等人[4]提出了一种基于特征学习的 K 均值描述符方法；向程谕等人[5]分别提取目标的 RGB 和深度特征，结合线性 SVM 进行分类。上述研究在早期的 RGB-D 图像识别研究中取得了一些成果，依靠先验知识的特征构建方法虽然能够在一定程度上改善 RGB-D 物体识别的精度，但是该类方式不利于进行非同类数据集的扩展，且精度的提高空间有限。

近年来，深度学习在图像处理的研究中逐渐呈现出它的优势[6]。因此，科研人员将 RGB 图像与深度图像相结合，并利用深度学习提升 RGB-D 物体识别的准确率，这种方法开始取代基于先验知识的特征获取方法，成为当前研究者们关注的热点。Socher 等人[7]提出单个卷积层与递归神经网络相结合的网络架构。殷云华等人[8]设计了一种将 CNN 与极限学习机相结合算法结构。但是浅层的网络结构并不能发挥深度学习的优势。Eitel 等人[9]提出了称为 colorjet 的深度图像处理方法，将深度图像编码为与 RGB 图像兼容的三通道图像，使用 5 个卷积层提取 RGB 特征与深度特征，通过全连接层组合两种模态的特征，该方法将 RGB-D 数据集的识别结果提升到了 91.3%。Aakerberg 等人[10]在 Eitel 的方法上进行了改进，提出了另外一种深度图像处理方法，并将网络层数提升到了 16 层。但是已有模型仅仅将 RGB 图像特征和深度图像特征进行简单的拼接，仍存在 RGB-D 图像有用信息缺失的可能。

为进一步提高三维目标识别精度，本文提出了一种基于深度神经网络的 RGB-D 物体识别算法(Re-CRNN)，将双流卷积神经网络与递归神经网络相结合，对 RGB 图像和深度图像进行端到端的训练；基于残差学习模型减小网络参数，计算深度图像每个像素点的表面法线向量，编码为三通道表示；在 CNN 网络顶层采用了一种新的特征融合方式，用以获得 RGB-D 融合特征，融合后的特征经过 GRU 递归神经网络生成特征序列；最后，在实验中对比了不同非线性激活函数在融合框架上的表现结果，在华盛顿 RGB-D 数据集中验证了本文算法的性能。

2 深度图像预处理

2.1 深度图像可视化编码

当前深度学习网络大多针对 RGB 图像，RGB 图像与深度图像的特征差异较大，使用深度学习网络训练深度图像依赖于深度图像编码[10-12]，将单通道深度图像编码为与 RGB 图像兼容的三通道表示，利用迁移学习的方法微调网络参数进行训练(如图 1 所示)。本文首先使用递归中值滤波[10]减少噪声干扰，重建缺失的深度值，对单通道深度信息每个像素点计算表面法线，将得到的表面法线归一化为单位向量，根据该点的空间坐标$[A_1,A_2,A_3]$被编码为 $A_1 \to \mathrm{R}$ 、 $A_2 \to \mathrm{G}$ 、 $A_3 \to \mathrm{B}$ 的三通道，并映射到整数值$[0, 255]$之间，增强深度图像的三维表达能力。在对单模态深度图像的训练中，编码后深度图像的识别结果比原始深度图像提高了 23%。

2.2 输入图像归一化

CNN 需要固定网络输入图像的尺寸，实现这个目标最简单的方法是将图像随机裁剪或缩放为正方形(如图 2(a)所示)，但是直接缩放或裁剪会损失图像原始比例，导致几何形变，如图 2(b)所示。

实验过程中发现这种忽略被识别物体的比例特征

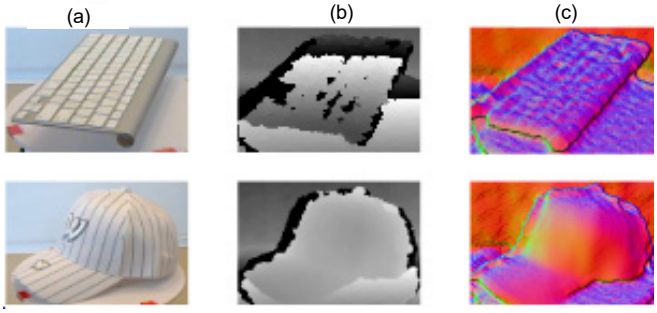


图 1 深度图像编码。(a) RGB 图像；(b) 原始深度图像；(c) 编码后深度图像

Fig. 1 Depth image encoding. (a) RGB image; (b) Original depth image; (c) Encoded depth image

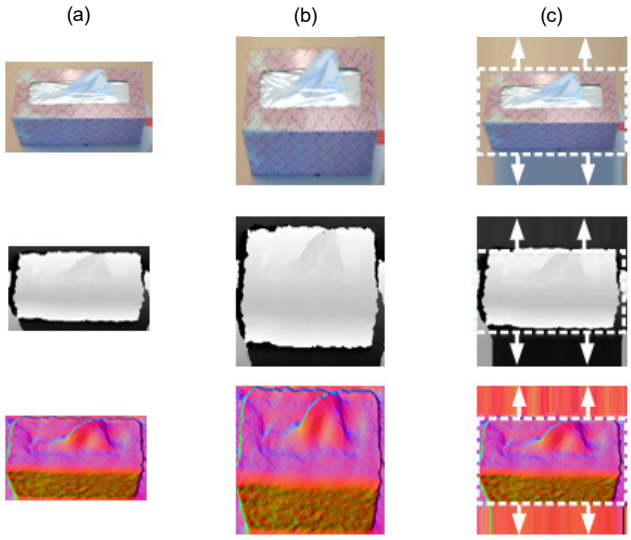


图 2 图像预处理。(a) 原始图像；(b) 直接缩放图像；(c) 短边扩充图像

Fig. 2 Input image preprocessing. (a) Original image; (b) Direct zoom image; (c) Short edge extended image

会降低物体的空间几何信息识别性能，与文献[9]中实验结论相同。所以，本文对样本图像进行归一化预处理：目标图像长边保留原始比例缩放为 256 pixels，短边按照长边缩放后的像素差值进行额外边界创建，并沿短边轴扩充获得 256 pixels×256 pixels 的图像，原始目标于扩展图像居中位置，如图 2(c)所示，白色的虚线框中保留图像的所有原始信息，框外为扩充的边界。

3 Re-CRNN 算法模型

本文算法结构主要由三部分组成：① 主干网络基于改进残差学习的双流卷积神经网络，每个数据流网络参数设置相同，分别对 RGB 图像和深度图像进行训练，提取高阶特征；② 一个新的特征融合单元，将 CNN 顶层的 RGB 特征和深度特征跨通道信息整合；③ GRU 递归神经网络。网络架构具体如图 3 所示。预训练阶段分别使用 ImageNet 数据集上的预训练权重来初始化 RGB 图像和深度图像，并根据华盛顿 RGB-D 数据集微调网络参数，生成 RGB 层和 Depth 层的参数模型。由于数据集中样本数量有限，对所有的图像进行旋转、缩放、随机裁剪进行数据增强，保留原始图像标签，丰富样本空间。

3.1 残差学习

在 RGB-D 图像的特征学习过程中，如果用两个数据流网络同时训练 RGB 图像和深度图像，则参数计算量较大。为了提高模型的计算速度，借鉴残差神经网络(ResNet50)[13]，基于残差学习对本文网络模型进行改进。ResNet 的思想是通过恒等映射(identity mapping)的跳跃式连接，将学习目标分解为多个求残差的过程，从而减小网络的学习参数，解决深层卷积神经网络的梯度弥散问题。原始残差学习网络第一层卷积核的大小为 7×7。为提高网络的复用率，使 3 个 3×3 的卷积核进行替换，卷积后的感受野与一个 7×7 的卷积核卷积的感受野大小相同，并降低了第一个卷积过程中的采样损耗，提取了更多的细节信息。本文中 RGB 与深

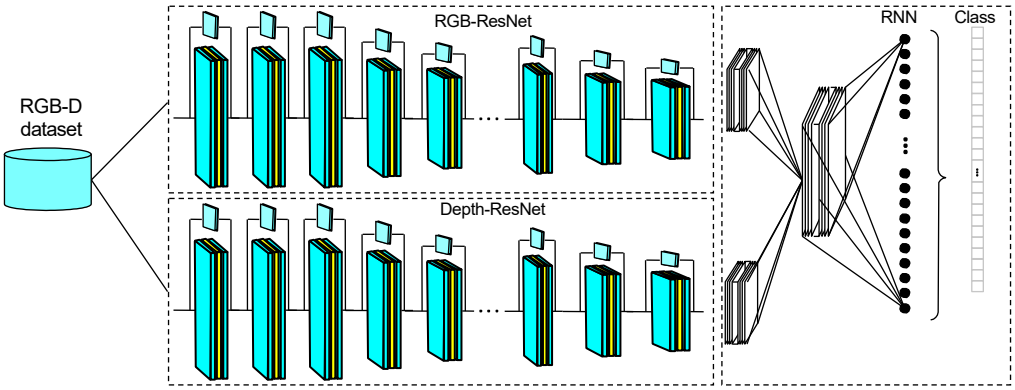


图 3 网络模型

Fig. 3 Network model

度两个数据流网络采取相同的方案。

假定网络层的输入为 s，经过中间层的期望输出为 $H(s)$，残差学习通过恒等映射跳过中间层将 s 作为初始输出，此时中间层需要学习的特征 $F(s)=H(s)-s$，学习目标不再是完整的 $H(s)$，而是一个残差 $H(s)-s$，通过多条支路使当前层的输入直接传输到更深的网络层，有效地减少了计算参数。相较于 VGG-16[11]的 153 亿次浮点运算，50 层的 ResNet 仅包含 38 亿次浮点运算，降低了网络的复杂程度。

残差单元表示为

$$Q_i = h(s_i) + F(s_i, W_i) \quad , \tag{1}$$

$$s_{i+1} = f(Q_i) \quad , \tag{2}$$

式中：s_i 表示第 i 个残差单元的输入，s_{i+1} 表示 s_i 的输出，即下一个残差单元的输入；F 是学习的残差，W_i 代表第 i 个残差单元的卷积操作，当 $h(s_i)=s_i$ 时表示恒等映射，f 代表激活函数。当 $h(s_i)=s_i$，$f(Q_i)=Q_i$ 时可以计算出层到深层 I 所学习到的目标特征：

$$s_I = s_i + \sum_{d=i}^{I-1} F(s_d, W_d) \quad 。 \tag{3}$$

通过链式求导计算出反向过程的梯度：

$$\frac{\partial f_{\text{loss}}}{\partial s_i} = \frac{\partial f_{\text{loss}}}{\partial s_i}\frac{\partial s_i}{\partial s_I} = \frac{\partial f_{\text{loss}}}{\partial s_I}\left(1 + \frac{\partial}{\partial s_i}\sum_{d=i}^{I-1} F(s_d, W_d)\right) \quad , \tag{4}$$

式中：第一个偏导 $\partial f_{\text{loss}} / \partial s_i$ 是 Loss 函数 f_{loss} 到 I 的梯度，另一项偏导代表通过权重层传播的梯度，常数 1 保证了快捷连接机制中梯度即使不断衰减也不会完全消失。

3.2 特征融合单元

为了避免已有的深度学习网络中：决策层获取每种模态的识别率[14]，问题注重单独模态的识别结果，忽略了 RGB 图像和深度图像的潜在互补特征；以及全连接层组合时，在一定程度上提高了识别结果，但是简单的拼接特征并不能利用两种模态的全部信息。所以在网络中构建了一个特征融合单元，如图 4 所示。将 ResNet 提取的高层次特征融合为新的 RGB-D 特征。去掉了 ResNet 的最后一个全连接层，在特征融合前将两个数据流网络中 conv5_x 输出的 feature map 合并起来，组合新的 fusion feature map，使用递归神经网络生成融合特征的高阶表示。4.2 节中的实验结果证明了本文方法优于文献[9]提出的全连接层融合的方式。

特征融合单元增加了一个 1×1 卷积层、一个批量

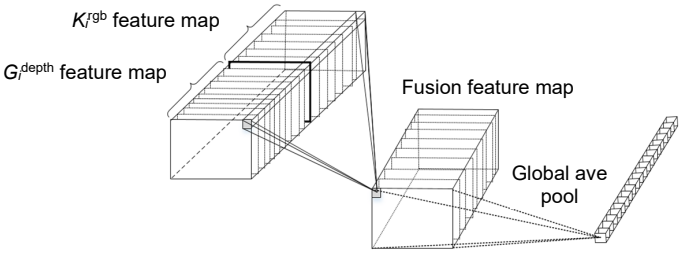


图 4 特征融合单元

Fig. 4 Feature fusion unit

归一化层、ReLU 激活函数和全局均值池化层。文献[15]使用 1×1 的卷积对不同层的特征图进行升维降维。融合层加入 1×1 的卷积对 RGB 特征和深度特征跨通道信息整合，并调整维度。用 $K^{\text{rgb}}=[K_1,K_2,\cdots,K_i]$ 和 $G^{\text{depth}}=[G_1,G_2,\cdots,G_i]$ 表示多模态网络的输入样本，i 是输入的样本标签，K_i 和 G_i 分别对应输入的 RGB 图像和深度图像。生成的 K_i^{rgb} feature map 和 G_i^{depth} feature map 充分融合为新的 fusion feature map。经过批量归一化和一个 ReLU 非线性激活函数提高网络泛化能力，对所有 feature map 进行全局均值池化，并将输出的结果排列起来，第 i 个标签对应的 RGB 图像和深度图像生成的融合特征被表示为

$$\boldsymbol{X}_i=[K_i^{\text{rgb}};G_i^{\text{depth}}] \quad 。 \tag{5}$$

3.3 递归序列

最近研究表明，卷积递归相结合在多模态深度学习中具有优势[16]。RNN 中先前序列的所有输入会共同作用当前序列的输出，卷积神经网络可以提取深层次的语义信息，RNN 重复利用 CNN 提取到的融合特征生成更好的特征表示，学习 RGB-D 图像中的潜在互补信息。传统递归神经网络在梯度传播的过程中先前输入序列的权重会逐渐减小，易出现梯度消失的问题。本文使用递归神经网络改进模型，被称为 GRU 递归神经网络[17]，GRU 递归神经网络通过两个门循环控制单元实现网络的长期记忆，更新门 Z_t(update gate)控制从先前隐藏状态到当前状态的信息，避免参数的丢失，其表达式：

$$Z_t=\alpha([\boldsymbol{M}_Z\boldsymbol{x}]_t+[\boldsymbol{U}_Z h_{t-1}]_t) \quad , \tag{6}$$

式中：α 为 sigmoid 激活函数，$\boldsymbol{x}$ 为时刻 t 输入的特征向量，h_{t-1} 为 $t-1$ 时刻的隐藏状态，$\boldsymbol{M}_Z$ 和 $\boldsymbol{U}_Z$ 是所学习的权重矩阵。

重置门(Reset gate) R_t 对融合特征进行过滤，减少冗余信息，增加鲁棒性，其表达式：

$$R_t=\alpha([\boldsymbol{M}_R\boldsymbol{x}]_t+[\boldsymbol{U}_R h_{t-1}]_t) \quad , \tag{7}$$

式中：$\boldsymbol{M}_R$ 和 $\boldsymbol{U}_R$ 对应于不同时刻的权重矩阵，候选状态 $\tilde{h}_t$ 与当前状态 h_t 分别表示为

$$\tilde{h}_t=z_t h_{t-1}+(1-z_t)h_t \quad , \tag{8}$$

$$h_t=\beta([\boldsymbol{M}\boldsymbol{x}]_t+[\boldsymbol{U}(R\odot h_{t-1})]_t) \quad , \tag{9}$$

式中 β 是 tanh 激活函数。重置门输出值较低时，遗忘先前的隐藏状态并使用当前输入复位，从而有效地忽略不相关的信息，生成更紧凑的特征表示。

4 实验及结果分析

验证实验使用两个公开的 RGB-D 数据集，在 Ubuntu16.04 操作系统下 6 核 i7-6700 CPU、单个 NVIDIA 1080GPU、8 根 16 G 内存条的深度学习工作站进行，并使用 Tensorflow 框架作为网络模型融合的基础。

4.1 数据集

1) RGB-D object dataset

华盛顿大学的 Lai 等人[1]公开的 RGB-D 对象数据集包含 300 个对象、51 个类别，总计约 250000 张 RGB-D 图像。验证实验中，每隔 5 帧对数据进行二次采样，生成 41877 幅 RGB 图像和对应的深度图像。随机抽取每个类别的一种对象用于测试，得到大约 35000 张训练图像和 7000 张测试图像。RGB-D 数据集中的部分样本如图 5 所示。

2) RGB-D sence dataset

本文在背景更复杂的 RGB-D 场景数据集[18]上验证本文算法对于不同数据集的有效性。该数据集包含 8 个不同的场景近 6000 张 RGB 图像和深度图像，所有的数据样本通过 Kinect 采集，该数据集的部分样本如图 6 所示。

4.2 模型分析

在 RGB-D object dataset 上通过模型变化调整网络

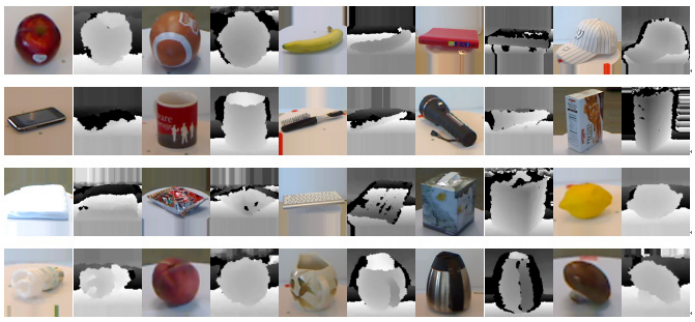


图 5 RGB-D 对象数据

Fig. 5 RGB-D object dataset

图 6 RGB-D 场景数据集

Fig. 6 RGB-D scene dataset

的最优性能。对比了不同非线性激活函数 tanh、elu、sigmoid、ReLU、softplus 对融合网络的影响，结果如图 7 所示。tanh 函数在零点梯度为 1，有利于网络中梯度的传播，提升模型的非线性表达能力；其次是 elu 激活函数，正区间的线性部分在一定程度缓解了梯度消失，负区间的软饱和性可以增加对于输入的变化和噪声的鲁棒性。对于 RGB 图像和 RGB-D 图像 tanh 激活函数得出了最好的结果，在深度数据上 elu 函数的表现最优，对于 RGB 模态和融合结果低于 tanh 函数，总体差异较小可以忽略不计，因此，本文使用 tanh 作为融合网络非线性激活函数。

本文使用后期融合的方式组合 RGB 图像和深度图像的高阶特征。为了验证后期融合在本文算法的有效性，分别将 RGB-ResNet 和 Depth-ResNet 的第二组到第五组(conv_2、conv_3、conv_4、conv_5)卷积特征作为特征融合单元的输入，得到不同层级融合的识别结果。图 8 所示为不同层级的融合结果对比，深层网络要比浅层网络对特征的抽象程度要高，conv5_x 层的特征进行融合准确率更高。表 1 中对比了全连接层融合(Fc-RGB-D)、特征融合层分类(Fu-RGB-D)与本文融合方式(Re-CRNN)的实验结果：conv5_x 生成的卷积特征作为特征融合单元的输入，特征融合层中 1×1 的卷积通道数为 256，分类器为 Softmax。与本文融合方法相比，全连接层简单的拼接融合方式遗漏了两种模态之间的信息交互，随着网络层数的增加识别率接近饱和。特征融合单元的跨模态的交互过程生成新的递归序列后，GRU 的循环过程进一步扩大了 RGB-D 图像的特征表达效果，准确率得到了进一步的提高。

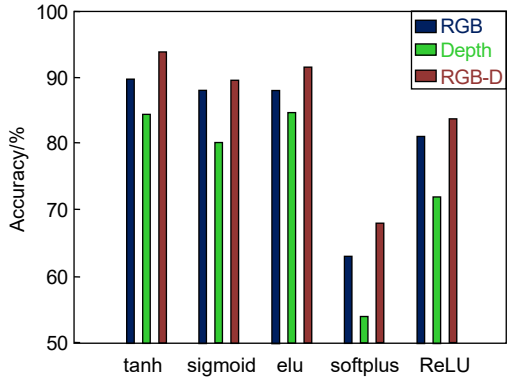


图 7 不同挤压函数对网络的影响

Fig. 7 Influence of different extrusion functions on the network

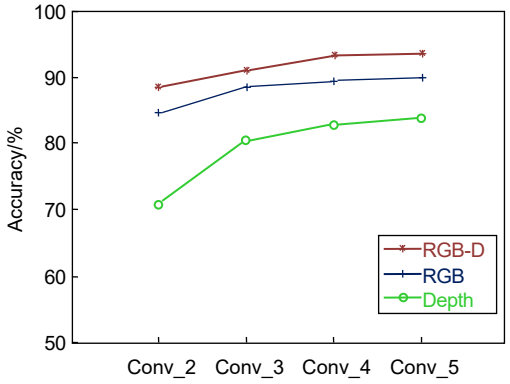


图 8 层级输出对比

Fig. 8 Level output contrast

表 1 特征融合方式对比

Table 1 Comparison of feature fusion methods

Method	Category accuracy/%	Instance accuracy/%
Fc-RGB-D+Softmax	93.2	96.8
Fu-RGB-D+Softmax	93.3	97.1
Re-CRNN	94.1	98.5

4.3 实验结果及分析

4.3.1 RGB-D 对象数据集

对象识别广泛处理两个不同的问题：实例识别和类别识别。实例(咖啡杯)代表独特的对象，而类别(杯子)代表共享相似特征(形状或结构)的实例[12]。按照 4.1 所述的 10 个随机分割进行了实验，训练前需对所有的深度图像进行可视化编码，然后将 RGB 图像和深度图保持原始比例缩放。实验设置预训练在 ResNet50 上获取 RGB 图像和深度图像的初始化模型，此阶段 RGB 图像和深度图像保持相同设置，初始学习率 0.0001，冲量 0.9，权重衰减 0.0001，批量 32，每种模态迭代 50000 次获得初始化模型，所需的时间为 2 h。融合网络使用 SGD 优化器训练我们的模型，取 10 次的平均值为实验结果，融合网络优化过程所需的时间约为 6 h。表 2 对比了其他算法在华盛顿 RGB-D 数据集上的分类结果。

表 2 中可以看出，RGB-D 图像的识别结果高于单独模态，证明融合 RGB 特征和深度特征可以进一步提高物体识别的准确率。本文提出 Re-CRNN 深度学习模型和融合方法在 RGB-D 数据集上获得了更好的识别结果，在类别识别中，RGB 图像的平均识别率为 90.3%，优于其他方法。深度图像的识别结果略低于文献[10]，结果相差较小，而 RGB 图像和融合后的 RGB-D 图像识别结果均表现出明显的优势，相较于文献[7]提出的 CNN-RNN 模型类别准确率提高了 7.3%。

图 9 展示了各类别分类结果的混淆矩阵(数据来源于第一个随机分割)，行的索引给出了 RGB-D 数据集中所有类别的真实标签，列的索引给出了各类别的预测结果，对角线的结果表示正确分类的总体占比，可以清晰地看到容易错分的对象。图 10 列举了总数据集中部分容易分错的样本，容易错分的对象存在于颜色和纹理都相似的物体，具体集中在以下几个类别：

表 2 与其他方法对比

Table 2 Compared with other methods

Method	Category accuracy/%			Instance accuracy/%		
	RGB	Depth	RGB-D	RGB	Depth	RGB-D
Bo et al[3]	82.4±3.1	81.2±2.3	87.5±2.9	92.1	51.7	92.8
CNN-RNN[7]	82.9±4.6	60.4±5.6	86.8±3.3	-	-	-
HCAE-ELM[8]	84.3±3.2	82.9±2.1	90.2±1.5	-	-	-
CNN-features[19]	83.1±2.0	-	89.4±1.3	92.0	45.5	94.1
Fus-CNN[9]	84.1±2.7	83.8±2.7	91.3±1.4	-	-	-
MM-LRF-ELM[11]	84.3±3.2	82.9±2.5	89.6±2.5	91.0	50.9	92.5
Andreas et al[10]	89.5±1.9	84.5±2.9	93.5±1.1	-	-	-
STEM-CaRFs[12]	88.8±2.0	80.8±2.1	92.2±1.3	97.0	56.3	97.6
Re-CRNN	**90.3±1.8**	**84.3±2.2**	**94.1±0.9**	**97.5**	**58.7**	**98.5**

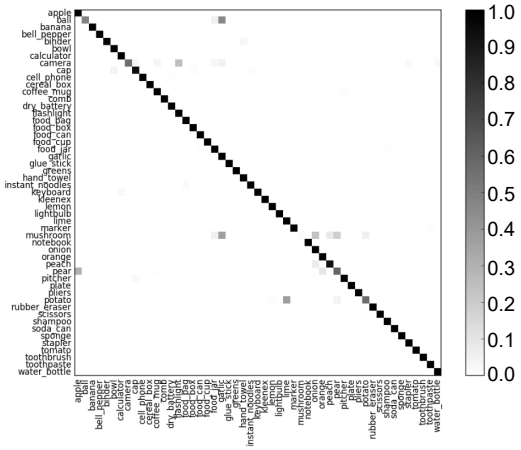


图 9 RGB-D 对象数据集的混淆矩阵

Fig. 9 Confusion matrix on RGB-D object dataset

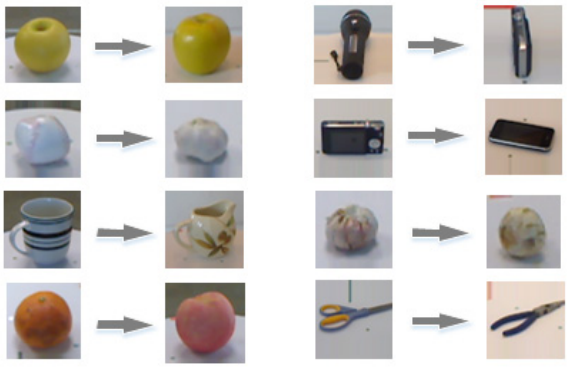


图 10 RGB-D 数据集中容易错分的对象

Fig. 10 Examples of misclassification in RGB-D object dataset

橙子类(orange)和桃子类(peach)、球类(ball)和大蒜类(garlic)、蘑菇类(mushroon)和大蒜类(garlic)等。

以上数据表明，首先，较少的实例会影响分类结果，如蘑菇类仅有 3 个实例，训练样本的单一化导致可学习特征类与量均受到限制，网络无法泛化新增添的数据，是造成错误的分类的原因之一；其次，具有 RGB 视觉分层上的相似数值的实例，在分类过程中较难被辨别，使结果产生偏差。深度图像的分类依据是物体的几何形态，上述对象高度的类间几何相似性使得深度图像的区分度降低，也会影响到最终的识别结果；此外，受传感器性能影响，已有的 RGB-D 数据集中图像的分辨率普遍不高，且深度图像中物体边缘部分深度值缺失，也可能对结果造成干扰。

4.3.2 RGB-D 场景数据集

室内场景识别是典型的多分类问题，场景图像更加密集地记录了场景中的所有物体。本文在 RGB-D 场景数据集上进行了额外实验来验证本文算法的普适性，对不同的室内场景进行分类，从该数据集中抽取 1434 张 RGB 图像和深度图像，共 8 类场景，每类场景中包含多种相似物体。将所有的图像尺寸调整为 256×256，并对深度图像完成表面法线编码。从每类场景中随机挑选 80%的样本用于训练，剩余的 20%用于测试。其它实验设置与 RGB-D 对象数据集相同，该数据集上的分类结果如表 3 所示。

实验结果显示，本文方法在 RGB-D 场景数据集上的实验结果比之前的结果提高了 4.5%，说明面对更加复杂的场景，深度学习的算法比手工设计特征描述子更有竞争性，Re-CRNN 可以在复杂的图像中有区别地提取 RGB 信息和深度信息，并且能够有效地完成 RGB 特征与深度特征的融合。实验结果同时表明，对于背景杂乱的场景分类问题，不同模态的信息互补是提高分类准确率的有效途径。场景数据集中分类结果的混淆矩阵如图 11 所示。

其中，纵坐标表示真实样本标签，横坐标表示预测标签。容易错分的场景主要存在于 desk_2 和 desk_3，table_small_1 和 table_small_2，分析其原因，在场景识别中最具有区别性的有用特征是不同的目标对象，根据这些对象的分布和包含的不同语义特征进行分类。而错分的场景中包含许多种相似的物体，如笔记本电脑、易拉罐、食品盒、杯子等，不仅具有相似的颜色信息，也包含相似的纹理信息，易产生检测数据的混淆，造成错误识别；深度图像采集时目标深度值相近且缺少明显的标志性数据特征，也会对识别结果造成一定影响，错分的场景示意如图 12 所示。

表 3　RGB-D 场景数据集分类结果

Table 3　RGB-D scene dataset classification result

Method	RGB/%	Depth/%	RGB-D/%
SIFT+Gis[5]	82.3	70.1	87.4
SIFT+PCA-Gist[20]	86.1	77.6	90.9
Re-CRNN	93.4	86.7	95.6

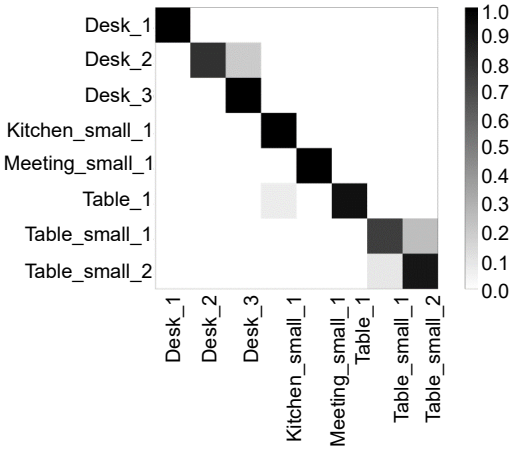


图 11　RGB-D 场景数据集的混淆矩阵

Fig. 11　Confusion matrix on RGB-D sence dataset

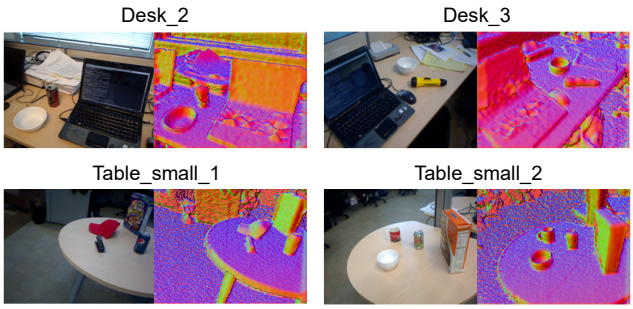


图 12　RGB-D 场景数据集错分实例

Fig. 12　Examples of misclassification in RGB-D sence dataset

5 结 论

本文提出了一种基于双流卷积递归神经网络的 RGB-D 物体识别算法 Re-CRNN，利用中值滤波去噪重建深度图像缺失的深度值，引入特征编码提高识别效果，数据样本通过数据增强获得扩充。使用两个并行的深度卷积神经网络对 RGB 图像和深度图像进行特征提取，基于残差学习的思想对模型效率进行提升。将 CNN 网络顶层提取的特征映射到一个公共空间，生成融合特征的高阶表示。最后在不同的数据集上与其他方法进行了实验对比，实验结果表明：RGB-D 图像比单模态 RGB 图像具有更好的识别效果，Re-CRNN 在华盛顿大学的 RGB-D 数据集识别准确率可达 94.1%。通过多个数据集的实验，证明本文算法具有较好的普适性。

参考文献

[1] Lai K, Bo L F, Ren X F, *et al*. A large-scale hierarchical multi-view RGB-D object dataset[C]//*Proceedings of 2011 IEEE International Conference on Robotics and Automation*, 2011: 1817–1824.

[2] Paulk D, Metsis V, McMurrough C, *et al*. A supervised learning approach for fast object recognition from RGB-D data[C]//*Proceedings of the 7th International Conference on PErvasive Technologies Related to Assistive Environments*, 2014: 5.

[3] Bo L F, Ren X F, Fox D. Depth kernel descriptors for object recognition[C]//*Proceedings of 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011: 821–826.

[4] Blum M, Springenberg J T, Wülfing J, *et al*. A learned feature descriptor for object recognition in RGB-D data[C]//*Proceedings of 2012 IEEE International Conference on Robotics and Automation*, 2012: 1298–1303.

[5] Xiang C Y. Research on feature extraction and classification method of RGB-D images[D]. Xiangtan: Xiangtan University, 2017: 28–31.
向程谕. RGB-D 图像的特征提取与分类方法研究[D]. 湘潭: 湘潭大学, 2017: 28–31.

[6] Li X, Li L P, Nan K K, *et al*. Face recognition method of smart home mobile robot[J]. *J Xi'an Poly Univ*, 2020, **34**(1): 61–66.
李珣, 李林鹏, 南恺恺, 等. 智能家居移动机器人的人脸识别方法[J]. 西安工程大学学报, 2020, **34**(1): 61–66.

[7] Socher R, Huval B, Bhat B, *et al*. Convolutional-recursive deep learning for 3D object classification[C]//*Proceedings of the 25th International Conference on Neural Information Processing Systems*, 2012: 665–673.

[8] Yin Y H, Li H F. RGB-D object recognition based on hybrid convolutional auto-encoder extreme learning machine[J]. *Infrared Laser Eng*, 2018, **47**(2): 0203008.
殷云华, 李会方. 基于混合卷积自编码极限学习机的 RGB-D 物体识别[J]. 红外与激光工程, 2018, **47**(2): 0203008.

[9] Eitel A, Springenberg J T, Spinello L, *et al*. Multimodal deep learning for robust RGB-D object recognition[C]//*Proceedings of 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems* (*IROS*), 2015: 681–687.

[10] Aakerberg A, Nasrollahi K, Rasmussen C B, *et al*. Depth value pre-processing for accurate transfer learning based RGB-D object recognition[C]//*Proceedings of the International Joint Conference on Computational Intelligence*, 2017: 121–128.

[11] Liu H P, Li F X, Xu X Y, *et al*. Multi-modal local receptive field extreme learning machine for object recognition[J]. *Neurocomputing*, 2018, **277**: 4–11.

[12] Asif U, Bennamoun M, Sohel F A. RGB-D object recognition and grasp detection using hierarchical cascaded forests[J]. *IEEE Trans Robot*, 2017, **33**(3): 547–564.

[13] Li L H, Wang Y X. Efficient 3D dense residual network and its application in human action recognition[J]. *Opto-Electron Eng*, 2020, **47**(2): 190139.
李梁华, 王永雄. 高效 3D 密集残差网络及其在人体行为识别中的应用[J]. 光电工程, 2020, **47**(2): 190139.

[14] Cheng Y H, Cai R, Zhao X, *et al*. Convolutional fisher kernels for RGB-D object recognition[C]//*Proceedings of* 2015 *International Conference on 3D Vision*, 2015: 135–143.

[15] Szegedy C, Liu W, Jia Y Q, *et al*. Going deeper with convolutions[C]//*Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition*, 2015: 1–9.

[16] Li X, Zhao Z F, Liu L, *et al*. An optimization model of multi-intersection signal control for trunk road under collaborative information[J]. *J Control Sci Eng*, 2017, **2017**: 2846987.

[17] Shen M Y, Yu P F, Wang R G, *et al*. Image super-resolution via multi-path recursive convolutional network[J]. *Opto-Electron Eng*, 2019, **46**(11): 180489.
沈明玉, 俞鹏飞, 汪荣贵, 等. 多路径递归网络结构的单帧图像超分辨率重建[J]. 光电工程, 2019, **46**(11): 180489.

[18] Ren X F, Bo L F, Fox D. RGB-(D) scene labeling: features and algorithms[C]//*Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012: 2759–2766.

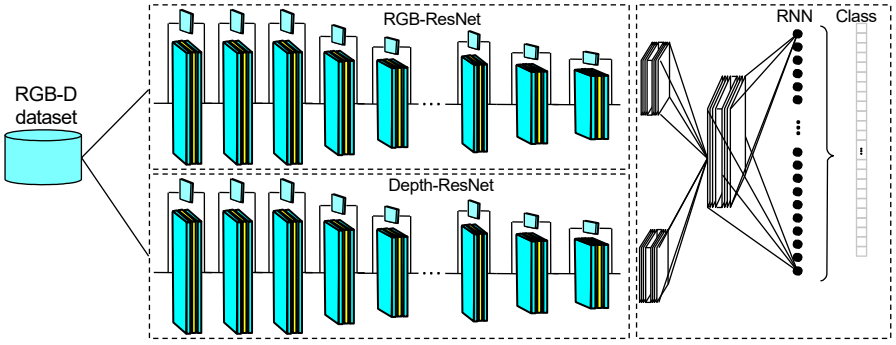
[19] Schwarz M, Schulz H, Behnke S. RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features[C]//*Proceedings of 2015 IEEE International Conference on Robotics and Automation* (ICRA), 2015: 1329–1335.

[20] Xiang C Y, Wang D L, Zhou Y, *et al*. Image classification based on RGB-D fusion feature[J]. *Comput Eng Appl*, 2018, **54**(8): 178–182, 254.
向程谕, 王冬丽, 周彦, 等. 基于 RGB-D 融合特征的图像分类[J]. 计算机工程与应用, 2018, **54**(8): 178–182, 254.

RGB-D object recognition algorithm based on improved double stream convolution recursive neural network

Li Xun[1,2], Li Linpeng[1*], Alexander Lazovik[2], Wang Wenjie[1], Wang Xiaohua[1]

[1]School of Electronics and Information, Xi'an Polytechnic University, Xi'an, Shaanxi 710048, China;
[2]Bernoulli Institute, University of Groningen, Groningen 9747 AG, Netherlands



Network model

Overview: The object recognition of RGB image is easily affected by the external environment, and the recognition accuracy has reached the bottleneck, which is difficult to meet the requirements of practical application. In recent years, the recognition method combined with depth image has become a new way to improve the accuracy of object recognition. The RGB image contains the color and texture features of the object, and the depth image contains the geometric features of the object and has illumination invariance. The fusion of RGB features and depth features can effectively improve the recognition accuracy. In order to make full use of the potential feature information of RGB-D image, and overcome the problem that the existing literature pays attention to the recognition results of single-mode and ignores the complementary advantages of RGB image and depth image, an RGB-D object recognition algorithm (Re-CRNN) based on improved double stream convolution recursive neural network is proposed. The depth image is encoded by calculating the surface normal. The depth image of a single channel is encoded into three channels. The transfer learning method is used to train the original image to generate the same level features as the RGB image. The backbone network is based on the double stream convolution neural network with improved residual learning. Residual learning is introduced to optimize the network structure and reduce the complexity of the model. The parameters of each data stream network are the same. The RGB image and depth image are trained respectively to extract the high-order features of RGB image and depth image. A feature fusion unit is added at the top layer of the network. The extracted high-level features of RGB image and depth image are fused across channels and mapped to a public space. Next, the fused features are further extracted by using a recursive neural network to generate a new feature sequence, which is classified by the softmax classifier. Finally, experiments are carried out on the standard RGB-D data set to compare the effects of different extrusion functions on the experimental results, as well as the fusion results of different convolution layers. The experimental results show that the recognition accuracy of RGB-D image is higher than that of RGB image, and the fusion of RGB features and depth features can further improve the accuracy of object recognition. The RGB-D object recognition algorithm proposed in this paper has achieved the best recognition results. The recognition accuracy rate on the RGB-D data set reaches 94.1%, which is obviously improved compared with the existing methods.

Li X, Li L P, Lazovik A, *et al.* RGB-D object recognition algorithm based on improved double stream convolution recursive neural network[J]. *Opto-Electron Eng*, 2021, **48**(2): 200069; DOI: 10.12086/oee.2021.200069

Foundation item: National Natural Science Foundation of China (61971339), Basic Research Program of Natural Science in Shaanxi Province (2019JM567), Science and Technology Guiding Project of China Textile Industry Federation (2018094), and Innovation and Entrepreneurship Training Programme for University Students (201910709019)
* E-mail: 771613990@qq.com