Journal of Image and Signal Processing 图像与信号处理, 2020, 9(4), 218-225
Published Online October 2020 in Hans. http://www.hanspub.org/journal/jisp
https://doi.org/10.12677/jisp.2020.94026

Hans汉斯

一种用于在架图书书脊语义分割的山字形网络

曾文雯[1]，杨　阳[2]，钟小品[2*]

[1]深圳大学图书馆，广东　深圳
[2]深圳大学机电与控制工程学院，广东　深圳
Email: *xzhong@szu.edu.cn

收稿日期：2020年9月23日；录用日期：2020年10月9日；发布日期：2020年10月16日

摘　要

在图像中识别在架书脊信息有助于实现更便捷的图书盘点，也可能实现即拿即走等更流畅的读者借阅体验，而书脊区域精确分割是重要前提。区别于普通目标分割，该分割问题的难点在于书脊的密集性及重复性。本文提出一种山字形深层神经网络结构，包含一个编码器及两个解码器。其中一个解码器为书脊分割主通道，另一个则结合书脊边界信息以融入更多的书脊边缘细节。另外，本文建立了一个书脊图像样本集，包含661张图像及15,454个手工标注的书脊实例。实验结果表明，提出的网络模型对书籍一类密集目标图像语义分割具有较高精度，在建立的样本集中具有约90%的均值交并比以及约95%的平均像素精度，性能优于经典的分割模型，验证了提出模型的有效性。

关键词

智慧图书馆，图书书脊，语义分割，深度神经网络

A Mountain-Shaped Network for Semantic Segmentation of Book Spines on-Shelves

Wenwen Zeng[1], Yang Yang[2], Xiaopin Zhong[2*]

[1]Shenzhen University Library, Shenzhen Guangdong
[2]College of Mechatronics and Control Engineering, Shenzhen University, Shenzhen Guangdong
Email: *xzhong@szu.edu.cn

Received: Sep. 23rd, 2020; accepted: Oct. 9th, 2020; published: Oct. 16th, 2020

Abstract

Identifying book spine on-shelves in the image can achieve a more convenient book inventory and

*通讯作者。

文章引用：曾文雯，杨阳，钟小品．一种用于在架图书书脊语义分割的山字形网络[J]．图像与信号处理，2020，9(4)：218-225. DOI: 10.12677/jisp.2020.94026

is possible to realize a better reader experience, such as take-and-go. Segmentation of the spine region is their important prerequisite. Different from ordinary target segmentation, the difficulty of this segmentation problem lies in that the spines are densely-packed and repeating. In this paper, a mountain-shaped deep neural network structure is proposed, which consists of one encoder and two decoders. One of the decoders is the main segmenting channel for the spine, and the other combines the spine interval information to incorporate more spine edge details. In addition, this research establishes a spine image sample dataset, including 661 images with 15,454 manually labeled polygons. The experimental results show that the proposed network model has high accuracy for semantic segmentation of dense target like book spine images, and has an average intersection ratio of 90% and an average pixel accuracy of 95% in the established dataset. The performance is better than the classical segmentation models, which verifies the effectiveness of the proposed model.

Keywords

Smart Library, Book Spine, Semantic Segmentation, Deep Neural Network

1. 引言

智慧图书馆建立在现代物联网技术基础之上，给我们带来更具魅力及亲和力的公共阅读、交流空间。新一代智慧型图书馆的发展将在多点突破，其中人工智能的快速发展为图书馆带来了新一轮转型的契机。

受零售行业无人商店思想的启发，实体图书馆一样可以实现无缝、流畅的借阅体验。其中的关键是通过机器视觉技术实现身份验证与物品识别。人脸识别技术为身份验证提供了成熟的方案，包括图书馆应用场景。而图书精确识别却是一个有待挖掘的领域，尤其是对于在架图书来说只有书脊部分可被观察到。为了实现便捷、准确的在架图书识别，本文将首先聚焦于图书书脊语义分割问题。如图 1 所示示例，每本图书均可被定位、分割出来以方便后续处理，其中绿色掩膜部分即为书脊区域。

(a) 原图　　(b) 理想分割结果

Figure 1. A segmentation example of book spine image
图 1. 书脊分割图例

图像分割是计算机视觉理解图像的重要一环，分割技术经历了从早期基于图的方法以及基于像素聚类的方法，到近期更为实用的基于深层神经网络的方法[1] [2]。2014 年 Long 等人[3]首次使用全卷积网络(Fully Convolution Networks)对自然图像实现端到端的像素分类框架，图像分割才真正产生了质的飞跃。全卷积网络使用了下采样会遗失细节信息，通过插值上采样将粗糙分割结果密集化，但上采样的热度图同样缺乏细节表达。Badrinarayanan 提出的 Segnet [4]以及 Ronneberger 提出的 U-net [5]则采用了编码解码结构的网络，是目前最为广泛采用的分割模型结构。Segnet、U-net 分别在道路场景和在医学图像场景中取得良好的效果，但对于非常稠密的目标，极易受到噪声的干扰，目标边界处常常无法正确分割。DeepLab [6]系列文章则使用空洞卷积，并在网络末端为每个像素构造条件随机场模型，以提高模型捕获细节的能力。笔者实践表明，对于相同书的不同实例，书脊间隙仍然有不少错分割区域，说明细节捕获能力仍有待改善。

如图 1 所示的在架图书场景则具有更多的挑战性，至少包含以下几点是没有被完全解决的。首先书脊目标特别稠密，目标位置的起止难以精准确定。其次是常有相邻多本一模一样的实例，传统分割算法会把他们合并起来当作一个实例。再次是异常摆放的书籍难以准确分割。

为了解决这些难点，本文提出一种编解码结构的多任务分割网络，称为山字形网络。其基本思想是在主通道分割目标的同时，在另一条通道上分割书脊的间隙，最后综合两个通道的结果即完成书脊语义分割。研究团队建立了一个包含 661 张图片样本数据集，每张图片以手工方式标注书脊区域，总共 15,454 个多边形区域。在该数据集上，实验结果表明，相对于传统的网络，提出的模型取得很好的效果，实现了均值交并比 90%、平均像素精度 95%的性能。

2. 山字形网络

受 SegNet 的编码－解码模型的启发，本文提出的山形网络结构如图 2 所示。该网络由一个编码器和两个解码器构成，是一个典型的多任务学习网络[7]。编码器是一个经典的卷积神经网络，由 10 个大小为 3 × 3 步长为 1 的卷积层组成，每个卷积层紧跟着一个校正线性激活单元(ReLU)。大小为 2 × 2 步长为 2 的最大池化层对特征图进行了下采样以减少特征规模，也使下一层特征具有更大的感受野。实际上，编码器取自经典 VGG19 网络的前十层，更多的卷积层需要更多的存储空间且已被证实对最终效果帮助甚微。两个解码器根据特征图分别重建书脊稠密特征及书籍间隙稠密特征。对这两个特征进行像素分类后得到书脊粗分割和间隙分割结果，书脊精确分割结果则由这些分割结果融合得到。

编码器典型的输入图像大小为 352 × 480，每次下采样尺寸减半，且在紧接着的卷积中对通道数加倍，最后得到 512 通道的 22 × 30 特征图。在解码器阶段，为了作上采样操作使特征图稠密化，网络在编码器对应位置使用转置卷积，同时接着的卷积层的特征通道数减半。为了保留更多的底层细节特征，使用了跳层连接，即每次转置卷积输出与编码器对应位置的特征图进行拼接。

书脊间隙非常小，我们期望间隙被归为背景一类。如果使用单任务学习，间隙信息会被学习过程抑制，这是因为间隙元素在前景中占比非常小。而把间隙单独作为一个分割任务有助于其得到强化。实验结果证实了这一点，间隙分割辅助任务可以改善书脊分割主任务的结果。

2.1. 书脊分割损失函数

书脊分割损失由 Softmax 层(如图 1)与交叉熵(cross-entropy)损失函数组合而成。对于书脊－背景这一二分类问题，标签 $y_i \in \{0,1\}$。交叉熵损失定义为

$$L_c = -\sum_{i=0}^{1} y_i \log p_i \tag{1}$$

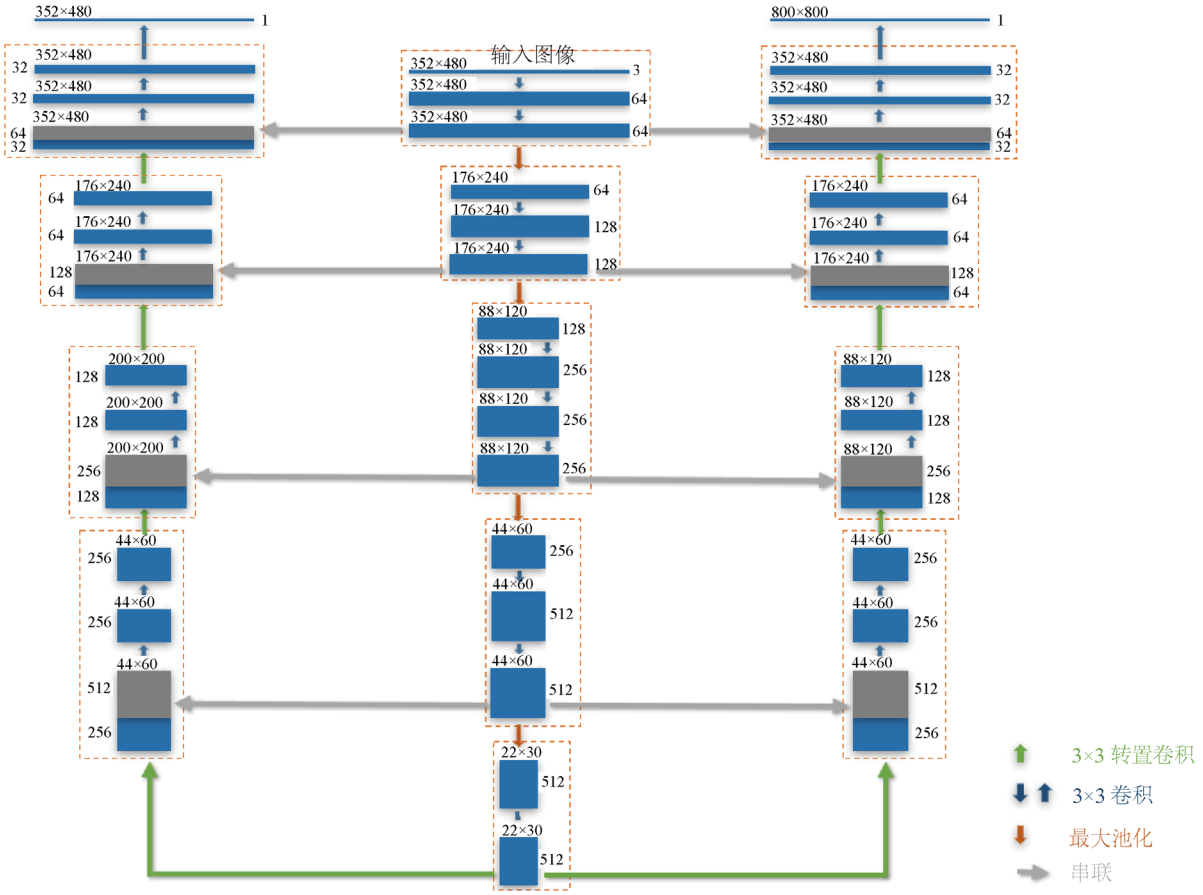


Figure 2. Mountain-shaped network structure
图 2. “山”字形网络结构图

其中 p_0、 p_1 分别是预测标签为 y_0 和 y_1 的 softmax 层归一化输出。

书脊区域与背景区域面积差别巨大，样本不平衡。为了较少占优标签样本对损失学习的贡献，使用参数 α_t 进行调节，则 $L_c = -\sum_{i=0}^{1} \alpha_i y_i \log p_i$ 。在本研究的实践中， $\alpha_0 = 0.4$ ， $\alpha_1 = 0.6$ 。

2.2. 书脊间隙分割任务

由于书脊间隙特别小，手工无法直接在样本上进行间隙标注。本研究所建立的数据集仅标注书脊部分，因此该数据集不可直接用于间隙分割任务。受文献[8]工作的启发，我们在书脊标注数据上应用图像处理等手段，获取书脊间隙的掩膜。具体流程如表 1 所示。

Table 1. Steps of obtaining the masks between the spines of books
表 1. 获取书脊间隙掩膜的步骤

输入书脊掩膜，输出书脊间隙掩膜
1. 使用霍夫变换等直线检测算法在书脊掩膜中提取书脊边缘直线； 2. 估计书脊边缘线上的书脊上下界； 3. 截取所有有效的书脊边缘线； 4. 每个有效书脊边缘线与书脊上下界所围多边形区域即为书脊间隙区域； 5. 集成所有的书脊间隙区域，形成书脊间隙掩膜。

由于间隙区域占比非常小，分割任务中正负样本比例极不平衡。传统的解决方案是按照类别比例增加权重或按类别比例采样，虽然容易分类的像素样本分割交叉熵损失小，但数量众多，对损失总贡献依旧大，产生了梯度被容易样本占优的问题。因此本文采用焦点损失(Focal loss [9])函数，重点学习不易分类样本，即

$$L_f = -\sum_{i=0}^{1} \alpha_i \left(1-p_i\right)^{\gamma} y_i \log\left(p_i\right) \tag{2}$$

其中为焦点参数，该参数作用是减小正确分类样本对损失的贡献。当 $\gamma = 0$ 时，上式退化为式(1)的交叉熵损失，随着 γ 的增加，对正确分类样本损失贡献的抑制力度增加。在本研究的实践中，设置 $\gamma = 2$ 。

3. 训练及测试结果分析

3.1. 样本获取及模型训练

本文的另一个重要工作在于建立了一个已经标注好的书脊图像样本库。在笔者所在高校资源丰富，具有藏书 323 万余册。本研究的样本来自于各种手机的拍照，包含不同的书籍种类(如中英文、文理科目等)、不同的拍照角度、不同的拍照时间(环境光不同)以及不同的拍照距离以覆盖各种可能的情况。采用 Labelme 平台进行手工多边形标注。目前数据集包含 661 个图像[1]，每个图像包含 20 到 30 个书脊实例，总共 15454 个手工标注实例，数据集规模还将随着研究的深入持续扩大。

本文所有训练及测试均采用 Matlab 环境及其深度学习工具箱。深度学习硬件配置主要有 Intel i7-4790 CPU，16G 内存以及 GTX1080 显卡。与本文提出方法进行对比的传统方法有五个，分别是 FCN16s、FCN32s、Segnet、U-net 以及 Deeplab-v3。输入图像尺寸统一为 352 × 480，使用数据增广技术增加样本的多样性，包括随机裁剪、旋转、颜色变化以及加入噪声等。

训练优化方法采用带动量的随机梯度下降法，动量因子设为 0.9。另外为了加速模型收敛，我们使用一个平滑递减的方式调整学习率，

$$lr = lr_0 \cdot \left(\frac{\max\_iter - iter}{\max\_iter}\right)^{0.9} \tag{3}$$

其中，lr_0 为初始学习率，$\max\_iter$ 是预设的最大迭代次数，$iter$ 表示当前迭代次数。

3.2. 分割结果及分析讨论

在相同的数据集上，本文提出方法的性能与主流的语义分割算法进行对比，如 FCN16s、FCN32s、Segnet、U-net、Deeplab-v3。一个分割示例如图 3 所示。我们分别采用均值交并比 mIoU 和平均准确度 mPA 来衡量分割性能。mIoU 定义如下，

$$mIoU = \frac{1}{2}\sum_{i=0}^{1} \frac{TP_i}{FP_i + FN_i + TP_i} \tag{4}$$

其中 TP_i、FP_i、FN_i 分别表示分类为 i 的真正、假正、假负的像素总数。而 mPA 定义如下

$$mPA = \frac{1}{2}\sum_{i=0}^{1} \frac{TP_i}{FP_i + TP_i} \tag{5}$$

全卷积神经网络将分类网络的全连接层替换为卷积层，池化下采样到最小分辨率后直接反卷积输出像素分类的热度图，这对于分割稠密的目标结果将显得粗糙。从图 xxx 不难发现，无论是全局补偿为 16

[1]标注数据集已开放下载，请参考链接 http://doi.org/10.4121/uuid:33f2a166-de13-4505-b359-2b202c491fd8。

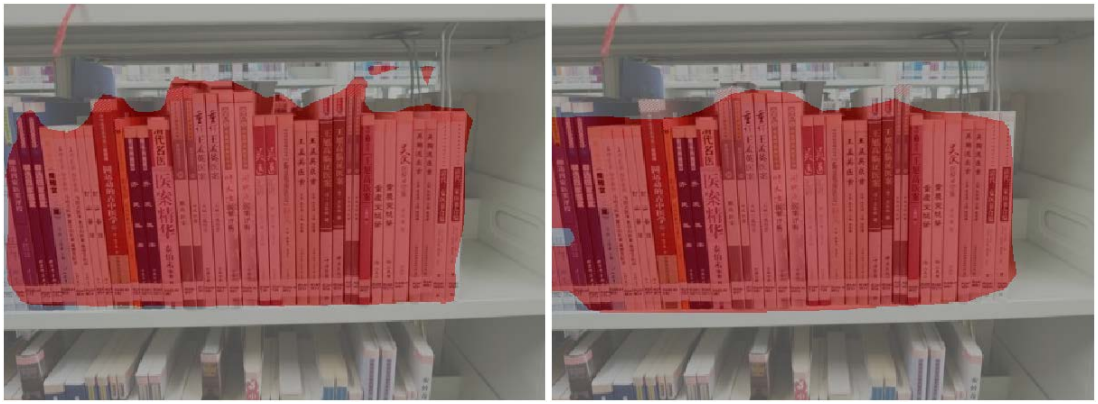
(FCN16)还是 32 (FCN32)，书脊区域光滑连成一片，书角处错分，也无法分割不同的书脊。

Segnet 和 U-net 则使用了编解码结构，在解码器上应用了多层小尺寸的反卷积，增加了计算量，但大大减小了参数量。该结构重视了分割的细节，书的边角可以正确分割，但书脊间隙仍然分割困难，且书脊不被当做整体看待，远处的背景以及不完整的书脊均有部分被当作目标。

Deeplab 采用了空洞卷积的方法，可以大大提高感受野，因此书脊被当作整体看待，远处背景以及不完整的书脊区域均可被正确分类。Deeplab 同时采用了残差神经网络的结构，网络总层数达到 100 层之多，参数量相对于其他网络也激增，如表 2 所示。从图 3 可以看出，无论是在书脊间隙还是在书脊边缘部分，提出模型相对于 Deeplab 方法均有较大的提升，且参数总量也大大优于 Deeplab。

Table 2. Performance comparison of algorithms
表 2. 算法性能比较

模型	可学习参数总量(M)	mIoU	mAP
FCN16s	134.3	0.8160	0.9063
FCN32s	134.3	0.8193	0.9072
SegNet	0.37	0.8660	0.8762
U-Net	31.0	0.7640	0.8750
deepLab v3	8479.1	0.8391	0.9186
提出模型	59.3	**0.8998**	**0.9514**



(a) FCN16s (b) FCN32s

(c) Segnet (d) U-net

(e) Deeplab v3　　(f) 提出模型

Figure 3. Sample results of common algorithms
图 3. 算法效果对比示例

由于样本不平衡的问题并不突出，书脊与背景的比例约为 0.4:0.6，表 2 显示的均值交并比和平均准确度数据差异并不大，它们对性能评价的效果是一致的，皆反映了提出模型比传统方法的分割性能有较大的提升，证实了提出模型的有效性。

表 2 的性能数据同时也说明了分割性能仍有较大的提高空间，一方面可以在模型研究上下功夫，另一方面数据集标注是由手工标记多边形完成，书脊处标注困难却对分割结果影响巨大，因此提高数据标注质量势在必行。由于在架的同一本书通常有多个副本，而索书号却是相同的。为了达成精准识别的目标，要求我们下一步对书脊进行实例分割。此外由于在架图书配置情况万千种，书脊信息有中文又用英文，也有多种不同的字体和字号，而笔者团队期望下一步建立的更大规模的数据集，训练好的模型泛化能力必定还有更大的提升空间。

4. 结论

为了实现更智能化的图书馆，利用计算机视觉技术对在架图书书脊进行分割是必经之路。书脊分割却是一个难点，因为在架书脊的密集性及相似性。本文因此建立了一个包含 661 张在架图书图像的数据集，并作手工标注。也提出了一种能解决密集书脊图像语义分割的“山”字形深度神经网络模型。该网络属于多任务学习网络，主分支针对书脊分割问题，另外还有一个书脊间隙的分割任务分支。实践结果证实，引入书脊间隙分割分支可以有效提升书脊分割的效果、提高分割精度。本研究的下一步计划建立体量更大、标注更精细的数据集，并实施实例分割以区分书籍的不同副本。

参考文献

[1] 田萱, 王亮, 丁琪. 基于深度学习的图像语义分割方法综述[J]. 软件学报, 2019, 30(2): 440-468.

[2] 张顺, 龚怡宏, 王进军. 深度卷积神经网络的发展及其在计算机视觉领域的应用[J]. 计算机学报, 2019, 42(3): 453-482.

[3] Long, J., Shelhamer, E. and Darrell, T. (2015) Fully Convolutional Networks for Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 7-12 June 2015, 3431-3440. https://doi.org/10.1109/CVPR.2015.7298965

[4] Badrinarayanan, V., Kendall, A. and Cipolla, R. (2017) Segnet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis Machine Intelligence*, **39**, 2481-2495. https://doi.org/10.1109/TPAMI.2016.2644615

[5] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-net: Convolutional Networks for Biomedical Image Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Munich, 5-9 October

2015, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

[6] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A.L. (2017) Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFS. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40**, 834-848. https://doi.org/10.1109/TPAMI.2017.2699184

[7] Ruder, S. (2017) An Overview of Multi-Task Learning in Deep Neural Networks. arXiv preprint, arXiv(1706): 05098.

[8] Zhou, X.Y., Zhuo, J.C. and Krahenbuhl, P. (2019) Bottom-Up Object Detection by Grouping Extreme and Center Points. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, 15-20 June 2019, 850-859. https://doi.org/10.1109/CVPR.2019.00094

[9] Lin, T.-Y., Goyal, P., Girshick, R., He, K.M. and Dollár, P. (2017) Focal Loss for Dense Object Detection. *Proceedings of the IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 2980-2988.