计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(12):238−242 [doi: 10.15888/j.cnki.csa.007164]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于改进 Faster R-CNN 的嘴部检测方法①

魏文韬, 刘　飞, 秦常程, 喻洪流, 倪　伟

(上海理工大学 康复工程与技术研究所, 上海 200093)
通讯作者: 魏文韬, E-mail: teyoutao@163.com

摘　要: 在通过嘴部进行人机交互的场景下, 外界光线变化、小目标检测的复杂性、检测方法的不通用性等因素给不同场景下嘴部的识别带来了很大困难. 该文以不同场景下的人脸图像为数据源, 提出了一种基于改进 Faster R-CNN 的人脸嘴部识别算法. 该方法在 Faster R-CNN 框架中结合多尺度特征图进行检测, 首先将同一卷积块不同卷积层输出的特征图结合, 然后对不同的卷积块按元素进行求和操作, 在输出的特征图上进行上采样得到高分辨率的表达能力更强的特征, 从而提高了嘴部这种小目标的检测性能. 在网络训练试验中运用多尺度训练和增加锚点数量增强网络检测不同尺寸目标的鲁棒性. 实验表明, 相比于原始的 Faster R-CNN, 对嘴部的检测准确率提高了 8%, 对环境的适应性更强.

关键词: 嘴部检测; Faster R-CNN; 多尺度特征; 卷积网络; 不同场景

引用格式: 魏文韬,刘飞,秦常程,喻洪流,倪伟.基于改进 Faster R-CNN 的嘴部检测方法.计算机系统应用,2019,28(12):238–242. http://www.c-s-a.org.cn/1003-3254/7164.html

Mouth Detection Method Based on Improved Faster R-CNN

WEI Wen-Tao, LIU Fei, QIN Chang-Cheng, YU Hong-Liu, NI Wei

(Institute of Rehabilitation Engineering and Technology, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: In the scenario of human-computer interaction by the mouth, the light changes, the complexity of the small target detection, and the detection method of none generality factors under different scenarios have brought great difficulties to detect the mouth. In this study, we take the face images with different scenarios as data source and propose a face recognition algorithm based on Faster R-CNN. In this method, multi-scale feature maps are combined in Faster R-CNN framework for detection. Firstly, we introduce a modified multi-scale feature map to effectively utilize multi-resolution information. Then, feature maps need to share the same size, so that element-wise sum operation can be performed. Features with higher resolution and stronger expression ability can be obtained by up-sampling on the output feature map. The detection performance of the small target is improved. In the training experiment, multi-scale training and increasing the number of anchor points are used to enhance the robustness of the network to detect targets of different sizes. Experiments show that the detection accuracy of the mouth is improved by 8%, and it is more adaptable to the environment compared with the original Faster R-CNN.

Key words: mouth detection; Faster R-CNN; multiscale feature; convolution network; different scenarios

① 基金项目: 上海地方能力建设项目 (16060502500)
Foundation item: Local Capacity Building Project of Shanghai Municipality (16060502500)
收稿时间: 2019-04-17; 修改时间: 2019-05-16; 采用时间: 2019-05-27; csa 在线出版时间: 2019-12-10

嘴部识别对于机器人视觉交互具有重要的研究价值. 给定任意一张人脸图像, 检测并确定嘴的位置, 在机器人控制交互式检测系统尤为重要. 实际场景下, 由于嘴部的姿态、脸部的表情和光线变化较大, 在不受约束的条件下拍摄图片, 高精度的嘴部检测是一个具有挑战性的问题. 在以往的方法中, 提取人工特征并将其作为二值分类进行建模求解已成为嘴部估计的标准步骤, 这种方法难以处理姿态各异、状态模型的嘴部. 因此, 建立一个机器人交互场景下的嘴态识别系统具有重要的理论与实际应用价值.

随着深度学习的发展, 各种目标检测算法被提出, 逐渐取代传统的检测算法. 卷积神经网络 (Convolutional Neural Networks, CNN) 是一种重要的深度学习方法, 它是一种前馈神经网络, 图像可以直接作为网络的输入, 可自动从图像数据中抽取特征, 避免传统识别算法中复杂特征提取和数据重建过程, 在图像识别领域应用广泛. CNN 在目标检测任务上表现出优越的性能[1–5], 其检测框架包括基于区域的方法与基于回归的方法[2]. 一类是 RCNN 系列的基于区域的目标检测算法[3], 如 Fast R-CNN、Faster R-CNN[4]以及 R-FCN 等, 这类算法的检测结果精度较高, 但速度较慢. 另一类是以 YOLO[2]为代表的将检测转化为回归问题求解, 如 YOLO、SSD[6]等, 这类算法检测速度较快, 但精度较低且对于小目标的检测效果不理想. 相比于候选区域的方法, 直接预测边界框的方法能提高目标检测系统的检测速度. 但 YOLO 网络直接对原始图像进行网格划分, 会使目标位置过于粗糙. SSD 加以改进, 对不同深度网络层回归采用不同尺度窗口, 因 SSD 采用的候选框选取机制, 对小目标的检测效果仍差于 Faster R-CNN.

Faster R-CNN 将区域生成网络 (Region Proposal Networks, RPN)[3]和 Fast R-CNN[4]检测网络融合, 实现了高精度的实时检测. 随着 Faster R-CNN 使用 CNN 网络结构层数由浅到深, 有 ZF[6]、VGG[7]、GoogleNet[8]和 ResNet[9]等, 尽管更深的网络可能带来更高的精度, 但会导致检测速度降低. 因此, 对于具体问题, 研究合适的基础网络结构和训练方法以保证较高精度的同时确保实时性, 是目前其主要研究方向之一[10].

为了解决复杂多变的交互场景下喂食机器人对于嘴的识别, 本文以喂食机器人与人交互这一任务为例, 基于 Faster R-CNN 目标检测网络进行改进实现人脸和嘴部的精确识别, 在 Caffe GPU 深度学习框架上进行实验. 结果表明, 采用改进的 Faster R-CNN 目标检测网络能够对人脸嘴部快速和精准的识别.

1 Faster R-CNN 简介

Faster R-CNN 是由 2 个模块组成: 生成候选区域的 RPN 模块和 Fast R-CNN 目标检测模块. RPN 模块产生候选区域, 并利用“注意力”机制, 让 Fast R-CNN 有方向性的检测目标. 首先, RPN 网络预先产生可能是人脸和嘴部的目标候选框, 然后 Fast RCNN 基于提取出的候选框来对目标检测识别.

1.1 区域建议网络

针对 R-CNN 和 Fast R-CNN 中 selective search 算法生成目标建议框的速度问题, Faster R-CNN 引入了区域建议网络代替 Selective Search 算法用于生成目标建议框[11], 极大地提升了目标建议框的生成速度.

RPN 的基本思想是在特征图上找到所有可能的目标候选区域, 它通过在原始的网络结构上添加卷积层和全连接层来同时在每个位置上回归目标边界框和预测目标分数. RPN 采用的是滑动窗口机制, 每个滑动窗口都会产生一个短的特征向量来输入到全连接层中进行位置和类别的预测. 在每个滑动窗口位置同时预测多个候选区域, 其中每个位置的预测候选区域的数量为 k. 因此, 回归层具有 $4k$ 个输出, 编码 k 个边界框的 4 个坐标, 分类器输出 $2k$ 个概率分数, 预测每个区域的所属目标的概率和所属背景的概率. “proposal”为目标生成层, 该层中剔除跨越边界的目标框, 并通过非极大值抑制[12]结合目标框前景得分筛选部分目标框, 最后通过目标框的回归信息得到 RPN 网络给出目标建议框, 最后选取 256 个目标建议框作为 RPN 网络的输出.

1.2 区域建议网络损失函数

在训练 RPN 网络时, 为每个候选框分配一个二值标签, 用于网络训练, 将以下 2 种情况分配正标签:

(1) 与某个真实目标区域框的 IoU (Intersection-over-Union) 最大的候选框.

(2) 与任意真实目标区域框的 IoU 大于 0.7 的候选框. 为所有真实目标候选框的 IoU 小于 0.3 的候选框分配负标签, 然后进行网络训练并微调参数. 图像的损失函数定义如式 (1) 所示.

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}\left(p_i, p_i^*\right) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}\left(t_i, t_i^*\right) \tag{1}$$

其中, i 表示小批次处理中的第 i 个候选框索引, p_i 是第 i 个候选框为目标的概率, 若 i 为候选目标, 则 p_i*

为 1, 否则为 0. $t_i=\{tx, ty, tw, th\}$是一个向量, 表示预测的参数化的候选框坐标. t_i*是对应的是真实目标框的坐标向量. t_i 和 t_i*的定义如式 (2) 所示.

$$\begin{cases} t_x=(x-x_a)/\omega_a, t_y=(y-y_a)/h_a \\ t_\omega=\log(w/w_a), t_h=\log(h/h_a) \\ t_x^*=(x^*-x_a)/w_a, t_y^*=(y^*-y_a)/h_a \\ t_w^*=\log(w^*/w_a), t_h^*=\log(h/h_a) \end{cases} \tag{2}$$

其中, (x, y) 为区域框的中心点坐标; (x_a, y_a)为候选框的中心点坐标; (x^*, y^*)为目标真实框的坐标, w 和 h 为包围框的宽和高. 算法的目的在于找到一种关系将原始框映射到与真实框 G 更接近的回归狂.

分类的损失函数L_{cls}定义如式 (3) 所示.

$$L_{cls}\left(p_i, p_i^*\right)=-\log\left[p_i^* p_i+\left(1-p_i^*\right)(1-p_i)\right] \tag{3}$$

回归的损失函数 L_{reg} 定义如式 (4) 所示.

$$L_{reg}\left(t_i, t_i^*\right)=R\left(t_i, t_i^*\right) \tag{4}$$

其中, R 是 $smooth_{L1}$ 函数, $smooth_{L1}$ 函数如式 (5) 所示.

$$smooth_{L1}(x)=\begin{cases} 0.5x^2, & |x|<1 \\ |x|-0.5, & |x|\geq 1 \end{cases} \tag{5}$$

1.3 Fast R-CNN

Fast R-CNN 负责对感兴趣区域进行类别分类和位置边框微调, 判断 RPN 找出的感兴趣区域是否包含目标以及该目标的类别, 并修正框的位置坐标. RPN 给出了 2000 个候选框, Fast R-CNN 网络需要在 2000 个候选框上继续进行分类和位置参数的回归.

首先挑选 128 个样本感兴趣区域, 使用 RoI Pooling 层将这些不同尺寸的区域全部下采样到同一个尺度上. RoI Pooling 是一种特殊的下采样操作, 给定一张图片的特征图, 假设该特征图的维度是 512×(H/16)×(W/16), 以及 128 个候选区域的坐标 (其维度为 128×4), RoI Pooling 层将候选区域的维度统一下采样成 512×7×7 的维度, 最终可得到维度为 128×512×7×7 的向量, 可将其看成是一批尺寸为 128、通道数为 512、大小为 7×7 的特征图. 此过程将挑选出的感兴趣区域全部下采样成 7×7 尺寸, 以实现权值共享. 所有感兴趣区域被下采样成 512×7×7 的特征图后, 以一维向量形式初始化前两层全连接层, 最后输入到用来分类的全连接层和边界框回归的全连接层.

2 Faster R-CNN 网络改进

2.1 多尺度特征图结合

Faster R-CNN 只利用最后一个卷积层的特征图进行目标检测, 无法更加精确的检测到一些更小的物体. 为了解决这一问题, 本文在 Faster R-CNN 的网络基础上结合多尺度特征图.

最近的许多研究表明了浅卷积层的特征图具有更高的分辨率, 有助于检测小目标. 这些方法表明, 结合不同卷积层的特征图可以提高检测性能. 本文在每个卷积块中利用多个层的特征图. 如图 1 所示, 首先将不同的卷积层连接到同一个卷积块中 (如 VGG-16 中的 conv5_3 和 conv5_2 层), 然后对不同卷积块的特征图 (如 VGG-16 中的 conv4 和 conv5 块) 进行元素求和. 因为不同的卷积块有不同大小的特征图, 需要共享相同的大小特征图, 这样才能执行元素的求和操作. 为此, 采用反卷积层放大后一层特征图的分辨率, 在原始模型中添加了一个 1×1 的步长为 2 的卷积层, 用于恢复特征图的大小, 因为经过上采样后特征图的大小比原始模型扩大了两倍, 实验证明改进后的多尺度特征图具有较高的精度和较低的计算成本.

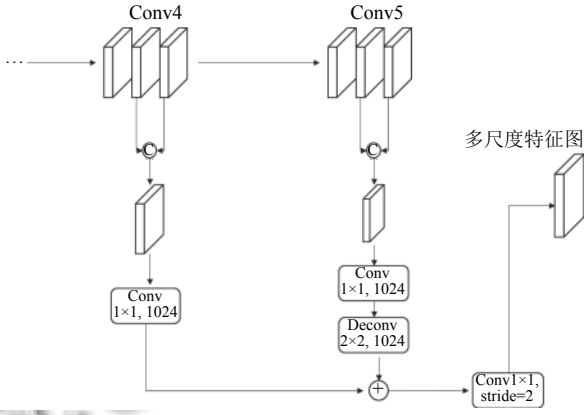

图 1 多尺度特征图结合

2.2 生成策略选择

不同尺度特征图生成策略对检测性能具有一定的影响. 一般来说, 集成更多的卷积特征图具有更高的检测精度, 但是会消耗更多的计算成本. 本文利用一种新的策略结合每个卷积块的多层特征图, 通过在原始 Faster R-CNN 中引入不同策略进行比较.

表 1 列出了在不同卷积层, 多尺度特征图上的 mAP 和处理时间. 从表 1 可见, 增加集成卷积特征图数量将提高网络的检测准确率. 然而, conv5_3+conv2, conv5_3+conv4_3+conv3_3 相对于 conv5_3+conv4_3 仅提高 0.1%, 且速度慢了一倍左右. 比较单个卷积层之间的结果, conv5_3/2 相对于 conv5_3 提高了 2.8% 的检测精度, 处理时间基本不变, 表明 conv5 中结合两

层特征图可以实现不同卷积层特征的互补, 提高了特征完整性. 同时 conv5_3/2+conv4_3/2 相对与 conv5_3/2 提高了 0.5% 的检测精度, 且仅增加了较少的额外计算成本. 最后, conv5_3/2+conv4_3/2 相对与 conv5_3+conv4_3 和 conv5_3+conv4_3+conv3_3 提高了检测精度的同时, 消耗了较少的额外计算成本. 上述结果表明, 改进后的生成策略比现有的策略更有效.

表 1 不同生成策略 mAP 和 FPS 对比

方法	mAP	推理时间 (FPS)
Conv5_3	73.2	15
Conv5_3+conv4_3	76.2	11
Conv5_3+conv4_3+conv3_3	76.3	6
Conv5_3/2	76.0	14
Conv5_3/2+conv4_3/2	**76.5**	**11**

3 实验

本实验主要为验证改进的 Faster R-CNN 算法在不同场景下的人脸图片上嘴部检测的有效性和优越性, 将该算法应用到机器人嘴部交互场景上, 能够让机器人快速完成嘴部的定位与检测, 完成相应的交互功能.

3.1 数据集

由于关于嘴部检测研究未曾发现公共数据集, 所以本实验使用的数据是自行采集和网上收集等方式获得的, 共 3000 张, 包含各种场景和不同质量的图片, 特别是光线较暗、成像质量较差、目标干扰、多角度的图像增多. 然后利用 LabelImg 工具对图像进行详细的标注. 根据实验要求, 将标注好的图像数据转换为 LMDB 格式. 如图 2, 对人脸和嘴部进行人工标注, 另外搜集了 1000 张不同场景和不同光线下的图片来进行方法测试, 对提出方法的有效性进行验证. 本研究使用 LabelImg 分别对训练集、验证集和测试集图片上的人脸和嘴部进行统一的标注.

3.2 网络训练

实验环境配置: GPU: GeForce GTX1050Ti, CUDA9.0, Ubuntu16.04, 显存 4 GB. 实验使用 caffe 深度学习框架进行相关代码和参数训练, 目标检测框架选择 VGG16 作为特征提取网络, 使用端到端的联合方式进行训练.

本文算法在训练网络模型的过程中, 为了能够使得梯度下降法有较好的性能, 需要把学习率的值设定在合适的范围内, 太大的学习率导致学习的不稳定, 太小值又会导致极长的训练时间. 训练模型的学习速率、衰减系数和动量参数的选取直接影响到最终的训练速度和结果, 本文选取一些较常用的学习率和衰减系数作为候选值, 如表 1 所示. 将衰减系数确定为 0.001, 学习速率的选取值有 0.1、0.01、0.001, 动量参数的选取值有 0.5 和 0.9, 在衰减系数不变的情况下, 首先确定了学习率, 然后确定动量大小. 其中当学习率为 0.1 时, 训练无法收敛, 可能是学习率初始值设置过大的原因. 由表 2 可知, 最终确定衰减系数为 0.1, 初始学习率为 0.001, 动量参数大小为 0.9. 如图 3 所示, 当训练迭代到 5×10^3 时, 损失函数值趋于平稳.

表 2 不同参数对应的测试精度

序号	衰减系数	学习速率	动量参数	mAP
1	0.001	0.1	-	/
2	0.001	0.01	-	0.9034
3	0.001	0.001	-	0.9128
4	0.001	0.001	0.5	0.9145
5	0.001	0.001	0.9	0.9343

由于 RPN 网络是 Faster R-CNN 和核心网络, 大大提高了获取候选框的效率, 由表 2 可知, Faster R-CNN 在不同的基础特征提取网络上的检测效果差异很大. ZF 网络相对于 VGG16 来说, 是一种小型的卷积网络, 将其作为 Faster R-CNN 的基础特征提取网络对人脸和嘴部检测识别 VGG16 网络的 mAP 基本能达到 90% 以上, 而 ZF 的 mAP 在 85% 左右, 但是 ZF 对图像的处理速度明显比 VGG16 大约快 3 倍. 在实际的交互场景下, VGG16 对每幅图像处理时间为 0.2 s 左右仍然是可以接受的, 因此, 综合考虑识别准确率和处理速率两个因素, VGG16 仍然优于 ZF 网络.

3.3 实验结果

图 2 是在不同的光线、角度和距离的场景下, 检测嘴部的效果, 结果表明本文的检测算法可以在光线较暗的场景下实现嘴部的定位和识别, 并且在一定的角度和距离场景下实现准确的检测. 准确率表明在机器人交互场景下, 算法的有效性和可靠性. Faster R-CNN 与本文算法比较如表 3 所示. 结果表明 Faster R-CNN 网络的准确率在 82.35% 左右, 测试耗时为 1.02 s 左右, 误检率为 12.32%, 漏检率 6.13%; 改进的 Faster R-CNN 网络的准确率在 92.43%, 测试耗时为 1.23 s 左右, 误检率为 4.58%, 漏检率为 3.28%. 由于是在 Faster R-CNN 网络上加入多尺度特征结合模块, 网络复杂度增加, 测试耗时较长, 将底层和高层的特征图进行融合, 对于不同尺度的目标能够准确的定位与识别. 对于小目标的检测率明显提高, 同时降低了误检率与漏检率.

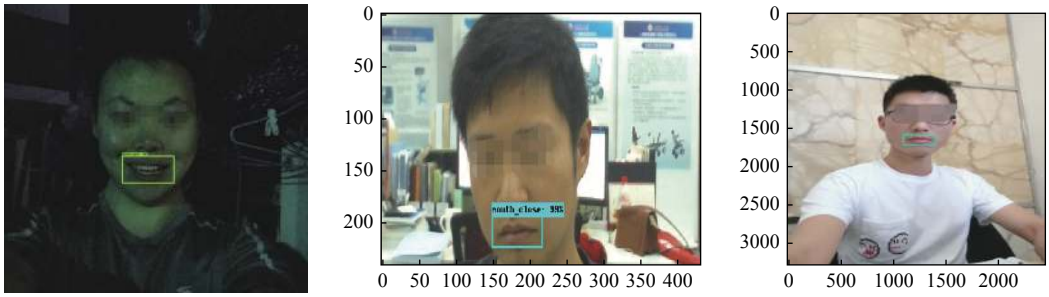

图 2 不同场景嘴部检测

表 3 检测结果对比

算法	测试耗时 (s)	mAP(%)	误检率 (%)	漏检率 (%)
Faster R-CNN	1.02	85.35	12.32	6.13
改进 Faster R-CNN	1.23	93.43	4.58	3.28

改进的 Faster R-CNN 是基于两阶段 R-CNN 框架, 其中检测是一个结合分类和边界框回归的多任务学习问题. 与目标识别不同, 需要一个交并比 (IOU) 阈值来定义正/负. 然而, 通常使用的阈值 u(通常 u=0.5) 对正样本的要求相当宽松. 产生的检测器经常产生噪声边界框 (FP). 假设大多数人会经常考虑相似正负样本, 通过 IOU≥0.5 测试. 虽然在 u=0.5 条件下获得的样本丰富多样, 但会使训练能够有效地区分相似正负样本的检测器变得困难, 造成检测仍然存在一定的误差.

4 结论与展望

目标检测作为计算机视觉领域的基本任务一直受到科研人员的关注, 目标检测方法的性能直接关系到高层领域的研究. 但通用目标检测方法在小目标检测上效果不佳, 专门为小目标检测设计的方法通用性差. 故本文改进 Faster R-CNN 并应用到嘴部识别中, 引入了多尺度特征结合, 结合不同卷积层特征图, 实现对小目标的准确识别. 通过对嘴部目标识别的对比实验, 验证了改进的算法对不同场景下的嘴部识别具有较好的效果. 改进的 Faster R-CNN 要求高质量的图片, 通过添加不同场景图像, 昏暗的灯光下, 质量差, 目标干扰的训练数据集可以有效提高目标识别的准确性目标检测算法在低像素和复杂环境和提高识别算法的鲁棒性. 接下来的研究工作是多网络进一步改进, 减少检测计算成本和时间, 对网络进一步压缩, 使得能够在嵌入式设备上完成实时检测任务.

参考文献

1 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521(7553): 436–444. [doi: 10.1038/nature14539]
2 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 779–788.
3 Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. arXiv: 1311.2524, 2013.
4 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(6): 1137–1149.
5 Girshick R. Fast R-CNN. IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 1440–1448.
6 Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B, *et al*, eds. Computer Vision-ECCV 2014. Cham: Springer, 2014, 8689: 818–833.
7 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv: 1409.1556. 2014. 1–14.
8 Szegedy C, Liu W, Jia YQ, *et al*. Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 1–9.
9 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 770–778.
10 Huang G, Liu Z, Van Der Maaten L, *et al*. Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. [doi: 10.1109/CVPR.2017.243]
11 Cai ZW, Vasconcelos N. Cascade R-CNN: Delving into high quality object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 6154–6162.
12 陈金辉, 叶西宁. 行人检测中非极大值抑制算法的改进. 华东理工大学学报 (自然科学版), 2015, 41(3): 371–378. [doi: 10.3969/j.issn.1006-3080.2015.03.015]