基于动作注意策略的树形DDQN目标候选区域提取方法

左国玉* 杜婷婷 马 蕾 卢佳豪 龚道雄
(北京工业大学信息学部 北京 100124)
(北京市计算智能与智能系统重点实验室 北京 100124)

摘 要：针对机器人在家庭环境下的目标检测问题，该文提出一种基于动作注意策略的树形双深度Q网络(TDDQN)目标候选区域提取的方法，该方法将双深度Q网络(DDQN)的方法与树结构的方法相结合，通过执行改变检测框的动作以使目标逐渐集中在检测框内。首先采用DDQN方法在执行较少的动作后选择出当前状态的最佳动作，获取符合条件的候选区域。然后根据执行所选择动作之后所得到的状态重复执行上述过程，以此构成树结构的多条“最佳”路径。最后采用非极大值抑制的方法从多个符合条件的候选区域选择出最佳候选区域。在Pascal VOC2007以及Pascal VOC2012上的实验结果表明，在不同数量的候选区域、不同阈值的IoU和不同大小以及不同种类对象的实验条件下，所提方法较其他方法都有着更好的检测性能，可以较好地实现目标检测。

关键词：目标检测；候选区域；树结构；双深度Q网络；动作注意
中图分类号：TP242 **文献标识码**：A **文章编号**：1009-5896(2019)03-0666-08
DOI: 10.11999/JEIT180358

Region Proposal Generation for Object Detection Using Tree-DDQN by Action Attention

ZUO Guoyu DU Tingting MA Lei LU Jiahao Gong Daoxiong
(*Faculty of Information Technology, Beijing University of Technology, Beijing* 100124, *China*)
(*Beijing Key Laboratory of Computing Intelligence and Intelligent System, Beijing* 100124, *China*)

Abstract: Considering the problem of object detection of robots in the home environments, a Tree-Double Deep Q Network (TDDQN) based on the attention action strategy is proposed to determine the locations of region proposals. It combines DDQN with hierarchical tree structure. First, DDQN is used to select the best action of current state and obtain the right region proposal with a few actions executed. According to the state obtained after executing the selected action, the above process is repeated to create multiple "best" paths of the hierarchical tree structure. The best region proposal is selected using non-maximum suppression on region proposals that meet the conditions. Experimental results on Pascal VOC2007 and Pascal VOC2012 show that the proposed method based on TDDQN has better detection performance than other methods for region proposals of different numbers, different Intersection-over-Union (IoU) values and objects of different sizes and kinds, respectively.

Key words: Object detection; Region proposal; Tree structure; Double Deep Q Network (DDQN); Action attention

1 引言

实现机器人对环境目标的检测、识别及定位一直是服务机器人要解决的关键技术，而目标检测是家庭环境下智能机器人技术研究的基础问题[1,2]。传统的检测方法是基于滑动窗口的区域选择策略[3]，但该方法没有针对性，时间复杂度高，窗口冗余，且人工设计的特征对于多样性的变化没有较强的鲁棒性。基于此，区域建议法利用图像中的纹理、边缘以及颜色等特征信息预先找出目标在图中可能会出现的位置，再对所提取到的候选区域进行分类。特征信息的使用可以做到在选取较少窗口的情况下仍能得到较高的召回率，降低了时间复杂度。常用

收稿日期：2018-04-18；改回日期：2018-11-16；网络出版：2018-12-04
*通信作者： 左国玉 zuoguoyu@bjut.edu.cn
基金项目：国家自然科学基金(61873008)，北京市自然科学基金(4182008)，北京工业大学智能制造领域大科研推进计划
Foundation Items: The National Natural Science Foundation of China (61873008), The Beijing Natural Science Foundation (4182008), The BJUT United Grand Scientific Research Program on Intelligent Manufacturing

的区域建议法有选择性搜索[4]和边界框回归[5]。后期，Girshick等人[6]将区域建议与卷积神经网络结合，设计了R-CNN框架，卷积神经网络的添加减少了人力，同时也削弱了检测窗口冗余的现象，目标检测性能取得了很大的进展。但是基于深度学习的检测方法从本质上讲，仍然是通过遍历图像以寻找与目标特征相似的区域，从区域选择的角度而言缺少了推理策略。

主动目标搜索策略[7]是另一种减少冗余窗口的方法，主要利用时间序列以顺序决策形式在视觉任务中收集先前时刻信息，并通过此信息决策滑动窗口位置。Caicedo等人[8]基于此提出主动目标定位法，采用深度Q网络(Deep Q Network, DQN)深度强化学习训练智能体(agent)，利用学习到的策略改变检测框大小和位置来选中目标。与主动目标定位法不同，Bellver等人[9]提出一种基于深度强化学习的分层目标检测方法，通过增加固定的层级表征来强制实现自上而下的搜索，将动作缩放到预期的感兴趣区域。这些方法都基于动作注意策略，但仅改变一个检测框的位置与大小，增加了找到理想候选区域的难度。文献[10]由此将树结构的强化学习用在了目标定位上，以提供多条搜索路径，即对多个检测框进行改变。但其中采用的DQN方法使用相同的最大Q值对动作选择和评估，很可能选择次优动作，以致需要执行更多动作才能检测到目标，影响最终检测结果。对此，Hasselt[11]针对Q学习存在的相同问题提出了双Q学习，并将改进策略用于DQN方法上，得到双深度Q网络(Double Deep Q Network, DDQN)方法，并在游戏领域取得了较好实验结果[12]。

针对以上工作，本文提出一种基于动作注意策略的Tree-DDQN (TDDQN)候选区域选择方法，通过执行改变检测框的动作以使目标逐渐集中于检测框内，进一步改进家庭环境下的目标检测效果。该方法将智能体的动作分为两组，每一组动作均使用DDQN选择当前状态的最佳动作，根据执行所选择动作之后所得到的状态再重复上一步骤，从而形成形似树结构的多条“最佳”路径。DDQN可选择最佳动作的特性使智能体在执行较少次数的动作下可以得到符合条件的候选区域，再结合树结构的多条路径特性得到多个这样的候选区域，更有利于选择出最佳的候选区域。

2　TDDQN候选区域提取方法

2.1　基于DDQN求取最佳动作

一般强化学习算法采取马尔可夫决策过程(Markov Decision Processes, MDP)作为数学模型，本文采用该模型进行区域提取。马尔可夫决策过程最主要的3个要素为动作、状态和奖励，其中：

动作A：代表一组有限的动作，有$a \in A$，其中的a_t表示第t时刻的动作。

状态S：代表一组有限的状态，有$s \in S$，其中的s_t表示第t时刻的状态。智能体在t时刻执行动作a_t后得到下一状态s_{t+1}，即$s_t \times a_t \to s_{t+1}$。

奖励R：对目标状态执行动作之后得到的立即奖励，即$S \times A \to R$。马尔可夫决策过程学习的是从环境状态到动作的映射，即$S \to A$，其具有延迟回报特点，因此可定义$Q = E(r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots)$来表示当前状态下动作策略的长期影响，其中$r_t$代表当前时刻的立即奖励，$\gamma$为折扣因素。通过不断学习，智能体逐渐学会在某一状态下选择何种动作更利于实现最终目的。

在本文方法中，智能体通过学习动作注意策略，针对当前检测窗口选择一系列动作不断改变其位置与大小，最终使目标物逐渐集中到检测框内。根据奖励R的定义，动作注意策略学习等价于训练智能体使其得到最大总的折扣奖励Q的一种策略。下文描述了基于动作注意策略的马尔可夫决策过程的实现过程。

2.1.1　动作

智能体执行的动作可以改变检测框的位置及大小，这里将智能体的动作分粗调和细调两组，其中粗调动作组用来将当前窗口缩放到子窗口中，粗略地对目标进行检测，可加快目标检测速度，细调动作组对当前窗口进行放大缩小及移动，对检测框进行微调，可提高目标检测的精准性。图1所示为两组动作示意图。

粗调动作组包含左上、左下、右上、右下、中间5个动作。智能体每一次对当前窗口执行粗调动作，所产生的子窗口大小将为当前窗口的0.6倍。细调动作组包含放大、缩小、左移、右移、上移、下移6个动作，智能体每次对当前窗口执行细调动作，所产生的子窗口大小为当前窗口的1.25倍(放大)或0.75倍(缩小)，或移动0.75倍的距离(左移、右移、上移、下移)。这里将动作分两组，除了以上作用外，还可与树结构结合形成多条搜索路径，以生成多个候选结果，再从中选择出最佳结果。

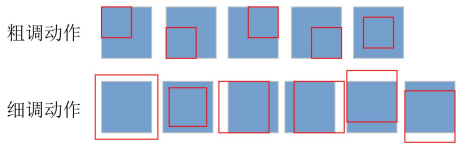


图 1 动作类型

2.1.2 状态

现有的对象提取算法主要是在不同位置和尺度上搜索可能的对象区域[5,13]，然而在搜索过程中，不同区域之间关键的相关性提示通常被忽略，如相对空间布局或语义相关性等，而这种情形通常与人类的感知过程相背离[14]，因为人类不会单独搜索每个局部图像中的对象，而是从感知整个场景开始，并通过顺序注意模式依次探索少数感兴趣区域。因此，本文区域提取模型的状态由当前窗口特征向量、整张图片的特征向量、所采取的历史动作3部分组成。鉴于vgg16具有很好的泛化能力，选取预训练好的vgg16模型对当前窗口特征向量和全局特征向量进行特征提取，以减少模型的训练时间。全局特征向量作为输入状态的一部分，为智能体提供上下文提示，便于智能体对当前检测框的状态做出决策，在加快检测速度的同时可提高检测精度。历史动作中的每个动作都由11维的2进制向量表示，除了所选择的动作对应位置为1外，其他都为0。

2.1.3 奖励

智能体在状态s下执行动作a后会收到一个奖励r，用于反映目标定位情况。该奖励函数公式设置为

$$r(s,a)=\begin{cases}+3, & \mathrm{IoU}(b',g)\geq\tau\\ \mathrm{sign}(\mathrm{IoU}(b',g)-\mathrm{IoU}(b,g)), & \text{其他}\end{cases} \tag{1}$$

其中，IoU(Intersection-over-Union)给出了目标物的真实检测框g和所观察区域的检测框b之间的相似性，可表示为$\mathrm{IoU}(b,g)=\mathrm{area}(b\cap g)/\mathrm{area}(b\cup g)$，$b'$是智能体执行动作后获得的新检测框，$\tau$是设定的一个阈值。

由式(1)可知，如果IoU大于等于τ，就认为暂且检测到目标，给予智能体奖励值+3，否则认为智能体没有检测到目标。如果没检测到目标，可使用2进制形式的$\mathrm{sign}(\mathrm{IoU}(b',g)-\mathrm{IoU}(b,g))$对智能体在当前检测框状态下所执行的动作是否有利于目标检测进行判断。如果智能体执行动作得到新检测框下的IoU大于执行动作之前的检测框下的IoU，则给予奖励值+1，否则–1。奖励函数的使用可以反映出哪些动作的执行，能让当前窗口更趋向于目标物的真实检测框，并进而加快智能体的学习速度。

2.1.4 DDQN求取最佳动作的实现

为解决DQN使用同一Q网络对动作选择与评估时出现次优Q解的问题，这里采用DDQN来提高选择最优动作的概率，即首先通过一个Q网络选择动作，再利用另一个Q网络评估动作，具体方法如下。

将提取的区域特征作为状态s_t，根据状态s_t执行某一动作得到下一状态s_{t+1}。动作选取采用$\varepsilon-$greedy策略，即智能体以ε概率分别从粗调和细调动作组中随机选择任意一个动作，以$1-\varepsilon$的概率分别从粗调和细调动作组中选择最好的动作。在状态s下执行动作a后产生结果的好与坏由最大的折扣奖励和决定，如式(2)

$$Q_m(s,a;\theta)=E(r_t+\gamma r_{t+1}+\gamma^2 r_{t+2}+\cdots|s_t=s,a_t=a) \tag{2}$$

其中，s为当前状态，a是当前状态下执行的动作，θ是网络参数，r_t代表当前时刻的立即奖励，γ是折扣因素。当$Q_m(s,a;\theta)$最优时，有

$$Q_m(s,a;\theta)=\max_a E(r_t+\gamma(r_{t+1}+\gamma r_{t+2}+\cdots|s_t=s,a_t=a)) \tag{3}$$

假设t+1时刻的状态是s_{t+1}，则式(3)可表达为式(4)

$$Q_m(s,a;\theta)=r_t+\gamma\max_a Q_m(s_{t+1},a_t;\theta) \tag{4}$$

利用式(4)进行对动作选择，即可得式(5)

$$a_t=\arg\max_a Q_m(s_{t+1},a_t;\theta) \tag{5}$$

在式(4)中增加一个Q网络，即可形成DDQN，如式(6)所示

$$Q(s,a;\theta')=r_t+\gamma Q(s_{t+1},\arg\max_a Q_m(s_{t+1},a_t;\theta);\theta') \tag{6}$$

式中，首先用最大的Q_m值选择一个最佳的动作，然后用另一个Q网络对此动作进行评估，再乘以折扣因子并加上立即奖励传给评估Q网络。

2.2 树结构路径生成

本节将使用树结构生成多条路径，其中每一条路径都采用DDQN方法，以此得到多条“最佳”路径。图2为树结构搜索过程，图中的实线表示所选择的粗调动作，虚线表示细调动作。以前两层为例，图中第1层的节点为当前状态，即候选区域。针对当前候选区域分别选择出在粗调动作数组和细调动作数组中预测值最高的两个动作；对检测框分别执行这两种动作，得到两个当前状态，即第2层中的两个节点所表示的候选区域。以此类推，可以分别得到树结构的第3层、第4层等等。

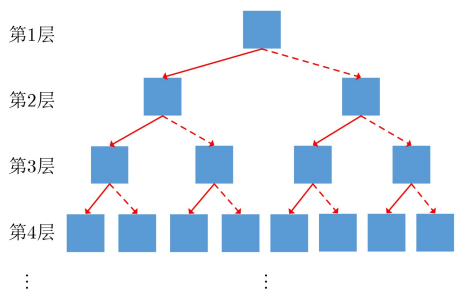
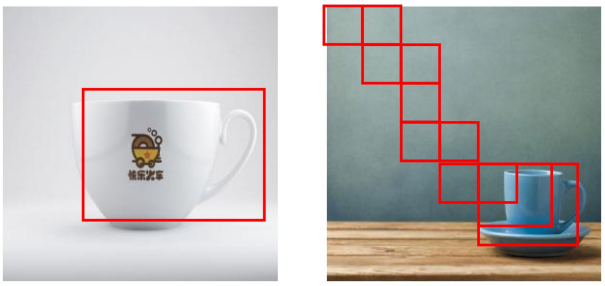


图 2 树状结构示意图

在对检测框执行动作的过程中，若所得的IoU大于τ，则停止搜索，这样可能出现图3所示的结果。图3(a)结果满足要求但并非最佳结果，如果检测框向左下角稍微移动就可得到更好的检测结果。针对这种情形，可终止此条路径，而树的其他路径可依旧继续执行，直到各自节点的IoU大于τ时为止，因此多条路径的选择可以增加检测到目标的概率。对于图3(b)所示情形，可以看到，很难在较少的层中检测到目标。一般情况下，这种执行动作数过多的路径不可能成为最终选择的理想路径，因此可对树的层数设置一个上限值n，这样无论节点是否达到IoU大于τ，只要层次数超过n，即可停止搜索。



(a) 检测不完整　　(b) 检测层数过多

图 3 可能出现的检测结果

根据以上分析，本文设计了如图4所示的树状搜索结构。假定图4中第3层右侧的两条分支出现IoU大于τ，则在第4层时停止搜索，因此只有第4层的左侧分支可以继续执行，以此类推，最后将得到的若干个检测框，采用非极大值抑制方法找到最佳的候选区域位置。

2.3 TDDQN候选区域提取方法实现

智能体每次执行动作完毕，会将当前的状态s_t、动作a_t、奖励r_t，以及下一时刻的状态s_{t+1}存入记忆向量$\boldsymbol{D}$中，当记忆向量积累到一定程度时，从中随机提取出一批样本进行训练。鉴于智能体在学习过程中采集到的样本是时间序列样本，样本之间具有连续性，因此，随机提取样本可以打破数据间的相关性，有助于模型的收敛。设随机提取出的样本下角标为j，则在训练阶段Q网络的损失函数可表示为$L=(Q_j-Q_m(s,a;\theta))^2$。图5为本文方法的网络结构，其中Q网络的更新采用了梯度下降法。

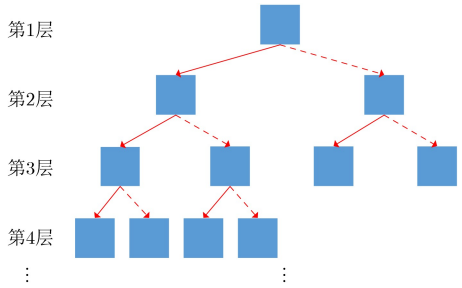


图 4 树结构生成示意图

首先使用vgg16网络提取图像特征，然后将当前状态输入到DDQN网络中，其中Q_m网络根据当前状态分别从两组动作中选择出最佳动作，由Q网络对所选择的动作进行评估。智能体执行完动作之后产生新的状态(候选区域)，并作为树结构的下一节点。再重复上述步骤，直到树结构的所有分支都到达检测标准或者树的层数上限时为止。最后对所产生的若干个候选区域采取非极大值抑制方法选择出最佳的候选区域。基于上述模型的候选区域提取方法的实现流程如表1所示。

3　实验与分析

3.1 实验设置

鉴于本文检测方法主要用于智能服务机器人对家庭环境场景的感知上，因此采用Pascal VOC2007, Pascal VOC2012数据集中与家庭环境相关的数据进行实验，代码运行在NVIDIA Tesla K40c GPU上。实验中，设置$\varepsilon-$greedy策略中的ε为1，并且ε以0.1的步长下降直到$\varepsilon=0.1$时为止。奖励函数中

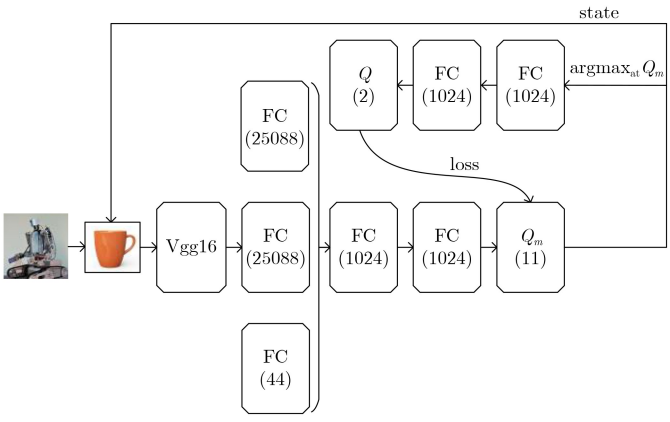


图 5 网络结构图

表 1 基于TDDQN的候选区域提取方法

输入	当前状态(候选区域，树的根节点)
输出	下一状态(新候选区域，树的子节点)
步骤1	初始化IoU的阈值τ和树的最大层次数n的值，并设树的初始层次数为1；
步骤2	根据当前状态，分别在粗调动作组和细调动作组中选择出经过DDQN方法得到的预测值最高的两个动作；
步骤3	执行粗调动作后得到的状态作为左节点，执行细调动作后得到的状态作为右节点；
步骤4	树的层次数加1；
步骤5	如果当前树的层次数小于n，并且仍有分支没有被截止，则执行步骤6，否则执行步骤7；
步骤6	如果左节点IoU大于τ，则予以截止，否则将左节点作为其所在路径的当前状态并执行步骤2；相应地，如果右节点IoU大于τ，则予以截止，否则将右节点作为其所在路径的当前状态并执行步骤2；
步骤7	对所有叶节点用非极大值抑制方法选取最优的候选区域。

的折扣因素γ为0.9。记忆容量为1000，批大小为100，学习率为1e-6，树层次数上限n为10。由于Pascal VOC2012没有公开基准真实数据标注(ground-truth annotations)，本文测试阶段召回率的比较只在Pascal VOC2007上进行，而其他指标在两个数据集上都做了对比实验。本文采用召回率R、准确率P和平均检测精度均值mAP 3个指标进行实验结果分析。召回率$R=\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$用来衡量算法被正确检测到的目标数量与真实目标数量的比例，准确率$P=\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$用来衡量被正确检测的目标数量与实际检测到的目标数量的比例，其中TP为正确检测到的目标数量，FN为正确目标被检测成负类的数量，FP为错误的目标被检测成正类的数量。平均检测精度均值mAP表示为

$$\mathrm{mAP}=\sum_{q=1}^{Q}\mathrm{AP}(q)/Q \tag{7}$$

其中，Q为检测目标的种类数，AP为某单个目标平均检测精度。

3.2 实验结果分析

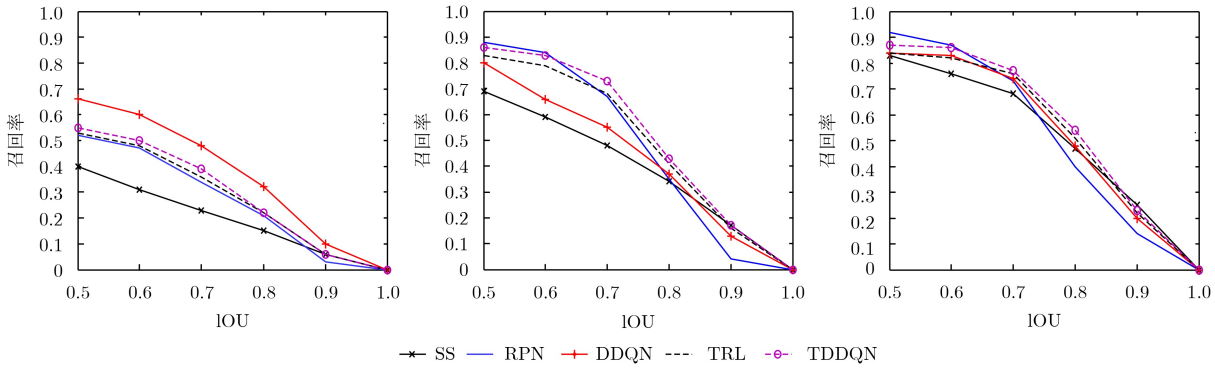
3.2.1 召回率

实验中将分别从候选区域数量、IoU值和目标检测物的大小对召回率的影响进行分析。下文图例中，SS, RPN, TRL和TDDQN分别表示选择性搜索(Selective Search)，区域建议网络(Region Proposal Network), Tree-RL和本文的TDDQN方法。

由于树最大深度为10，因此候选区域数量的最大值为1023，此处取候选区域数量的上限值为31, 255, 1023。考虑到目标物大小等因素，通常情况下目标在树达到5～8层时就能被给予最佳检测结果，因此这里可设31为较少的候选区域数量，1023为较多的候选区域数量，255为正常所需要的候选区域数量。图6为采用以上5种方法，IoU分别为0.5, 0.6, 0.7, 0.8, 0.9, 1.0时所得到的召回率曲线。

由图6(a)可知，当提取出的候选区域数量较少时，DDQN, TRL以及TDDQN方法具有强化学习中的推理性质，其较传统的SS方法具有更高的召回率。而TDDQN和TRL中由于存在多条搜索路径，在候选区域数量较少的情况下，其性能优势暂未体现出来，然而随着更多的候选区域的生成，如图6(b)所示，TDDQN和TRL方法逐渐显示出其优势。又由于TDDQN方法相比TRL增加了DDQN，因此会得到比TRL方法更好的候选区域。当提取出的候选区域数量足够多时，如图6(c)所示，各种方法产生符合条件的候选框的可能性相近，但从实验结果来看，本文方法的召回率要更好一些。与目前主流的RPN网络比较，当候选区域数量较少时，TDDQN优于RPN，随着候选区域数量的逐渐增加，在IoU为0.5和0.6时，TDDQN召回率略低于RPN，但是当IoU大于0.7以后，TDDQN得到的召回率都要大于RPN，因此平均来看，本文的TDDQN方法可以比RPN方法得到更高的召回率。

从IoU阈值对召回率的影响来看，当IoU阈值



(a) 31个候选区域 (b) 255个候选区域 (c) 1023个候选区域

图 6 不同数量候选区域在不同IoU下的召回率

设置为0.5时，检测要求较低，因此各种方法都表现出较好的性能。随着IoU阈值增加，检测要求越来越高，对所选择的候选区域的要求也相应越来越高，因此各方法的召回率随之降低。

实验把所占区域为整幅图像1/8的目标物划为较大目标物，否则为较小目标物。由于较小目标物相对较大目标物有着更小的区域面积，因此检测时需要更多的粗调与细调动作来提取出最佳的候选区域，这样会出现树深度超过10层的可能性，从而导致检测性能降低。为验证TDDQN的有效性，在IoU=0.5、候选区域数量上限为1023的情况下，采用不同方法进行实验，得到如图7所示的召回率。可以看到，无论目标物大与小，TDDQN都比强化学习中的其他方法有着更高的召回率，特别是针对较小目标物，检测效果更明显。

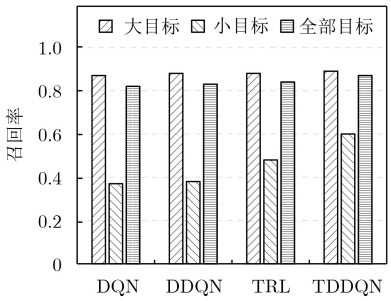


图 7 检测物大小不同时不同检测方法的召回率

3.2.2 检测精度

为进一步考察检测性能，采用RPN, DQN, DDQN, TRL, TDDQN分别训练同一检测器Fast R-CNN，对多种检测对象的平均检测精度进行实验比较。所有Fast R-CNN 检测器都经过在ImageNet上预训练好的ResNet-101网络微调得到，具体地，将ResNet-101网络最后的平均池化层与全连接层替换为一个新的全连接层，用来进行边界框回归和分类。实验也将当前流行的Faster R-CNN方法加入比较队列中。

表2和表3分别列出了Pascal VOC2007和VOC2012数据集中不同检测对象的实验结果。由表中数据可以看到，对比RPN与Faster R-CNN, TDDQN具有强化学习的推理策略，针对某一状态有着更高的决策能力，因此有着较好的检测性能。而针对强化学习的其他方法而言，本文将TRL与DDQN相结合，既具有TRL多条路径搜索的优势，又具有DDQN得到最佳Q值的优势，因此即便对于检测效果最差的对象，较其他方法也有着较好的结果。另外，从平均检测精度均值指标来看，无论在Pascal VOC2007还是VOC2012数据集下，各检测器的优越性依次为TDDQN+Fast R-CNN, TRL+Fast R-CNN, Faster R-CNN, RPN+Fast R-CNN, DDQN+Fast R-CNN, DQN+Fast R-CNN，因此进一步说明了本文方法有着更好的检测性能。

3.2.3 运行时间分析

运算速度是衡量目标检测算法性能的有效指标，实验取前节中检测性能前3的方法对单张图像检测消耗时间进行对比实验，结果如表4所示。可以看出，本文方法的检测效率虽略低于Faster R-CNN，

表 2　不同方法下Pascal VOC2007数据集中各检测对象的平均检测精度(%)

方法	瓶子	椅子	桌子	狗	人	沙发	电视机	平均检测精度均值
RPN (vgg16)+Fast R-CNN (ResNet-101)	54.3	60.2	70.8	84.1	76.2	78.7	73.0	71.0
Faster R-CNN (ResNet-101)	55.6	56.4	69.1	88.0	77.8	79.5	71.7	71.2
DQN (vgg16)+Fast R-CNN (ResNet-101)	50.4	54.3	61.8	80.2	71.1	73.5	68.9	65.7
DDQN (vgg16)+Fast R-CNN (ResNet-101)	52.6	55.2	61.3	80.5	71.3	74.0	69.1	66.3
TRL (vgg16)+Fast R-CNN (ResNet-101)	55.0	60.1	73.3	84.5	76.3	79.6	73.4	71.7
TDDQN (vgg16)+Fast R-CNN (ResNet-101)	55.7	60.2	74.2	85.3	77.4	79.6	73.7	72.3

表 3　不同方法下Pascal VOC2012数据集中各检测对象的平均检测精度(%)

方法	瓶子	椅子	桌子	狗	人	沙发	电视机	平均检测精度均值
RPN (vgg16)+Fast R-CNN (ResNet-101)	50.5	48.6	57.1	90.0	79.0	66.1	65.9	65.3
Faster R-CNN (ResNet-101)	50.8	48.5	59.0	91.9	80.5	66.3	65.4	66.1
DQN (vgg16)+Fast R-CNN (ResNet-101)	49.3	45.7	50.8	82.8	73.9	59.9	63.6	60.9
DDQN (vgg16)+Fast R-CNN (ResNet-101)	51.5	47.6	52.3	82.9	75.2	61.1	63.8	62.1
TRL (vgg16)+Fast R-CNN (ResNet-101)	53.1	51.7	55.6	87.8	80.7	66.6	67.6	66.2
TDDQN (vgg16)+Fast R-CNN (ResNet-101)	53.4	51.9	58.7	88.0	80.9	66.8	67.9	66.8

表 4 不同数据集上检测单张图片消耗的平均时间(s)

数据集	TDDQN (vgg16)+Fast R-CNN (ResNet-101)	TRL (vgg16)+Fast R-CNN (ResNet-101)	Faster R-CNN (ResNet-101)
VOC2007	0.9	1.6	0.4
VOC2012	1.0	1.8	0.5

但是3.2.2节中给出的检测平均精度却高于Faster R-CNN。相较于TRL方法，TDDQN+Fast R-CNN与TRL+Fast R-CNN两种方法都采用了Fast R-CNN检测器，因此，对单张图像检测消耗时间主要取决于于候选区域提取阶段。可以看到，由于本文方法在TRL的基础上融合DDQN方法，相当于对TRL方法做了优化，因此检测效率得到了很大的提升，消耗的时间也就更少。

3.2.4 算法收敛性分析

图8所示为TRL方法与TDDQN的收敛性能。可以看出，TRL方法大约需要300个回合数收敛，而本文方法大约250个回合数就可以收敛，因此本文方法较TRL方法有着更快的收敛速度。另外，图8中显示，TDDQN的均方误差相对更小，说明本文方法的策略精确度更高，进一步说明本文方法中DDQN存在的有效性好。

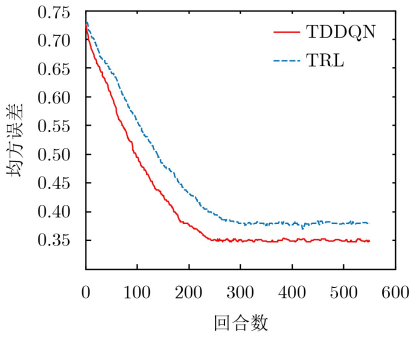


图 8 不同算法的均方误差曲线

3.2.5 可视化过程

图9示例了当树的层数为4时分别以瓶子、桌子、人为检测对象的可视化过程。图中，红色框为提取的候选区域，蓝色为所提取候选区域中IoU大于0.5的区域。

可以看出，当树层数为4时，第1排这样较小的目标物在较少的候选区域下的检测还存在困难，这也例证了为什么随着候选区域的不断生成，前文中图6(b)相较图6(a)能有更高的召回率。但是即便是在较少的候选区域数量下，从最后两排可以看出，本文方法对中上等大小目标的检测是可行的。图中也可看出，中间一排在第3层时已经提取出IoU大于0.5的区域，但是本文方法仅仅终止了此条路径，而其他搜索路径依然在进行搜索，因此在第4层时又提取出了另外IoU大于0.5的区域，通过对多个符合条件的结果进行比较，更有利于找到最佳检测结果。

通过前面的实验分析和图9可视化过程可知，相比较滑动窗口的固定路径搜索目标，本文方法可以针对不同对象以不同的路径进行搜索。与区域建议方法不同，本文采用的是高层的推理策略而不是低层的提示方法；与边界框回归算法相比，本文方法采用的不是单一的结构化预测方法，而是提出一

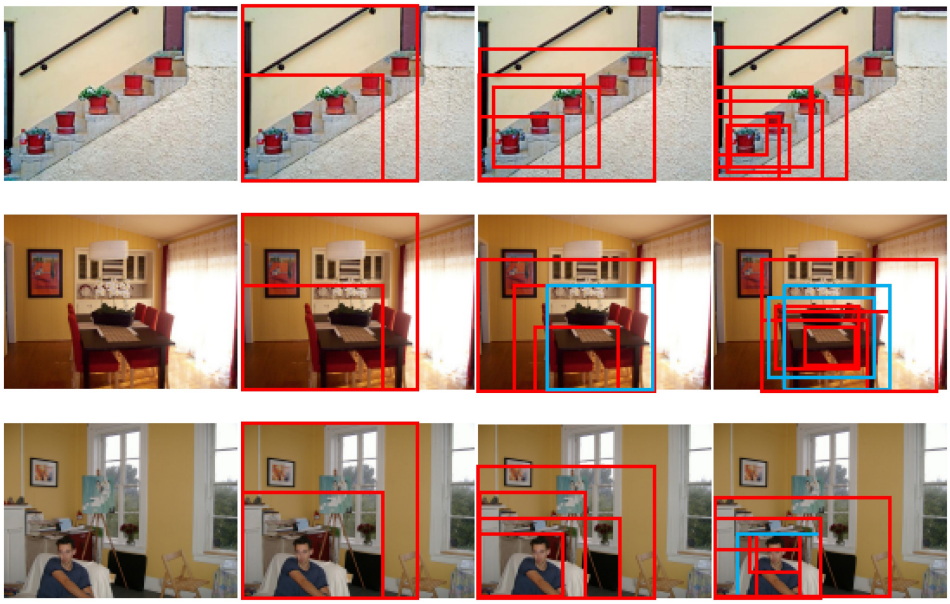


图 9 可视化过程

种动态的基于动作注意的策略，这种方法需要智能体针对当前区域的状态，执行改变检测框的动作以使目标越来越集中到检测框内，并针对检测框覆盖目标的程度给予奖励以促进智能体的行为。

4　结论

本文针对机器人在家庭环境下的目标检测方法进行了研究，提出了一种基于动作注意策略的TDDQN的目标候选区域提取的方法。智能体通过学习动作注意策略，针对当前检测窗口选择一系列动作不断改变其位置与大小，最终使目标物逐渐集中到检测框内。该方法将DDQN方法与树结构方法相结合，既能够通过得到的最佳Q值选择出最优的动作，在执行更少的动作数下得到符合条件的候选区域，又能够在多条路径下执行上一步骤以得到多个符合条件的候选区域，从多个符合条件的候选区域中选择出最佳的检测区域，以此提高检测的性能。实验表明，无论是在不同数量的候选区域、不同阈值的IoU、不同大小的目标对象，还是不同种类的目标对象下，本文方法较其他方法都有着较好的检测性能。

参 考 文 献

[1] TANG K, JOULIN A, LI L J, *et al*. Co-localization in real-world images[C]. Computer Vision and Pattern Recognition, Columbus, USA, 2014: 1464–1471. doi: 10.1109/CVPR.2014.190.

[2] 伍锡如, 黄国明, 孙立宁. 基于深度学习的工业分拣机器人快速视觉识别与定位算法[J]. 机器人, 2016, 38(6): 711–719. doi: 10.13973/j.cnki.robot.2016.0711.
WU Xiru, HUANG Guoming, and SUN Lining. Fast visual identification and location algorithm for industrial sorting robots based on deep learning[J]. *Robot*, 2016, 38(6): 711–719. doi: 10.13973/j.cnki.robot.2016.0711.

[3] DALAL N and TRIGGS B. Histograms of oriented gradients for human detection[C]. Computer Vision and Pattern Recognition, San Diego, USA, 2005: 886–893. doi: 10.1109/CVPR.2005.177.

[4] SANDE K E A V D, UIJLINGS J R R, GEVERS T, *et al*. Segmentation as selective search for object recognition[C]. International Conference on Computer Vision, Barcelona, Spain, 2011, 1879–1886. doi: 10.1109/ICCV.2011.6126456.

[5] ZITNICK C L and DOLLAR P. Edge boxes: Locating object proposals from edges[C]. European Conference on Computer Vision, Zurich, Switzerland, 2014, 391–405. doi: 10.1007/978-3-319-10602-1_26.

[6] GIRSHICK R, DONAHUE J, DARRELL T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Computer Vision and Pattern Recognition, Columbus, USA, 2014, 580–587. doi: 10.1109/CVPR.2014.81.

[7] GONZALEZ-GARCIA A, Vezhnevets A, and FERRARI V. An active search strategy for efficient object class detection[C]. Computer Vision and Pattern Recognition, Boston, USA, 2015, 3022–3031. doi: 10.1109/CVPR.2015.7298921.

[8] CAICEDO J C and LAZEBNIK S. Active object localization with deep reinforcement learning[C]. International Conference on Computer Vision, Santiago, Chile, 2015, 2488–2496. doi: 10.1109/ICCV.2015.286.

[9] BELLVER M, GIROINIETO X, MARQUES F, *et al*. Hierarchical object detection with deep reinforcement learning[OL]. http://arxiv.org/abs/1611.03718v2, 2016. doi: 10.3233/978-1-61499-822-8-164.

[10] JIE Zequn, LIANG Xiaodan, FENG Jiashi, *et al*. Tree-structured reinforcement learning for sequential object localization[C]. International Conference on Neural Information Processing Systems, Barcelona, Spain, 2016, 127–135.

[11] HASSELT H V. Double Q-learning[C]. International Conference on Neural Information Processing Systems, Whistler, Canada, 2010, 2613–2621.

[12] HASSELT H V, GUEZ A, and SILVER D. Deep reinforcement learning with double Q-learning[C]. Association for the Advancement of Artificial Intelligence, Phoenix, USA, 2016, 2094–2100.

[13] REN Shaoqing, HE Kaiming, 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, 39(6): 1137–1149. doi: 10.1109/TPAMI.2016.2577031.

[14] NAJEMNIK J and GEISLER W S. Optimal eye movement strategies in visual search[J]. *American Journal of Ophthalmology*, 2005, 139(6): 1152–1153. doi: 10.1038/nature03390.

左国玉：男，1971年生，副教授，研究方向为机器人学习与控制、人机交互.

杜婷婷：女，1991年生，硕士生，研究方向为强化学习、目标检测.

马　蕾：女，1992年生，硕士生，研究方向为深度学习、目标检测.

卢佳豪：男，1994年生，硕士生，研究方向为机器人学习、强化学习.

龚道雄：男，1968年生，副教授，研究方向为模式识别、计算智能.