(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 106373160 A
(43)申请公布日 2017.02.01

(21)申请号 201610797462.X

(22)申请日 2016.08.31

(71)申请人 清华大学
地址 100084 北京市海淀区清华园1号

(72)发明人 刘华平 张辉 孙富春

(74)专利代理机构 北京清亦华知识产权代理事务所(普通合伙) 11201
代理人 廖元秋

(51)Int.Cl.
G06T 7/70(2017.01)
G06N 3/04(2006.01)

权利要求书4页 说明书6页

(54)发明名称

一种基于深度强化学习的摄像机主动目标定位方法

(57)摘要

本发明提供了一种在图像采集应用中摄像机主动调整进行目标定位的方法，属于模式识别技术领域和摄像机主动定位技术领域。该方法包括训练一个评价摄像机定位效果的深度神经网络；进行多次目标定位试验，在定位实验过程中，训练一个拟合强化学习值函数的深度神经网络，通过深度神经网络判断摄像机“上转”、“下转”、“左转”、“右转”、“放大”、“缩小”和“不变”7种操作的优劣；采用决策网络根据摄像机当前获取的图像信息，对摄像机操作做出决策。该发明中提出的方法，基于深度强化学习算法，提高了采集图像的质量。能够适应不同的目标定位任务，自主学习定位方法，人为参与环节很少，是一个摄像机主动学习，自主目标定位的方法。

CN 106373160 A

1.一种基于深度强化学习的摄像机主动目标定位方法，其特征在于，该方法包括以下步骤：

(1)训练一个评价摄像机定位效果的深度神经网络，将该网络命名为评价网络N_R由多层神经网络组成；

(2)进行多次目标定位试验，在定位实验过程中，训练一个拟合强化学习值函数的深度神经网络，将该网络命名为决策网络N_Q，通过决策网络N_Q判断摄像机7种操作的优劣；

(3)完成决策网络N_Q训练后，采用决策网络N_Q根据摄像机当前获取的图像信息，对摄像机操作做出决策。

2.如权利要求1所述方法，其特征在于，所述步骤(1)具体步骤如下：

(1-1)设置评价网络N_R：评价网络N_R的网络结构依次为：输入层为RGB图像，图像高为H_{net}，宽为W_{net}，由于RGB图像为3个维度，所以输入层的维度为$H_{net}\times W_{net}\times 3$；$L_{RC}$层为卷积神经网络，激励函数为ReLU函数；$L_{RP}$层为全连接层，前$L_{RP}$-1层的激励函数也为ReLU函数，最后一层全连接层激励函数为Sigmoid函数，且设置维度为1，作为评价网络N_R输出，评价网络N_R输出定义为评价值；将评价网络N_R中的所有参数统一表示为θ_R，评价网络N_R逐层的运算过程表示一个函数映射，命名为评价函数 $F_{N_R}(\bullet|\theta_R):R^{H_{net}\times W_{net}\times 3}\rightarrow R$，其中$\bullet$表示网络的输入图像，实际计算中会输入不同的图像；R表示实数，函数意义为将维度为$H_{net}\times W_{net}\times 3$的实数空间图像映射到1维实数空间的评价值；

(1-2)采集训练评价网络N_R的数据集：

(1-3)从数据集D中随机挑选$M_{R,b}$组样本，记为批量样本$D_{batch}=\{d^1,d^2,\cdots,d^{M_{batch}}\}$，以及标签S中与其对应的批量标签$S_{batch}=\{s^1,s^2,\cdots,s^{M_{R,b}}\}$；

(1-4)根据步骤(1-1)的评价函数计算评价网络N_R对批量样本D_{batch}的评价值$F_{N_R}(d^i|\theta_R)$，其中$i=1,2,\cdots,M_{R,b}$；

(1-5)定义评价网络N_R的优化目标为$J_R=\frac{1}{M_{R,b}}\sum_{i=1}^{M_{R,b}}(F_{N_R}(d^i|\theta_R)-s^i)^2$，其中 $F_{N_R}(d^i|\theta_R)$为第i个样本d^i输入网络后输出的评价值，将最小化评价值和标签之间的欧式距离作为优化目标训练网络，计算优化目标对评价网络参数θ_R的梯度为$\frac{\partial J_R}{\partial \theta_R}$；

(1-6)采用随机梯度下降法，更新评价网络参数$\theta_R:=\theta_R-\alpha_R\frac{\partial J_R}{\partial \theta_R}$，其中$\alpha_R$为评价网络的学习率；

(1-7)重复上述步骤(1-3)～(1-6)，不断更新评价网络参数θ_R，直到评价网络收敛，完成评价网络训练，评价网络收敛的依据是连续C_R次优化目标J小于阈值η_R。

3.如权利要求2所述方法，其特征在于，所述步骤(2)具体步骤如下：

(2-1)设置决策网络N_Q结构依次为：输入层为RGB图像，图像高为H_{net}，宽为W_{net}，与评价网络相同；L_{QC}层为卷积神经网络，激励函数为ReLU函数；L_{QP}层为全连接层，前L_{QP}-1层的激励函数也为ReLU函数，最后一层全连接层无激励函数，设置维度为7，作为网络输出，将决策网络

N_Q中的全部参数统一表示为θ_Q，决策网络N_Q逐层的运算过程表示为一个函数映射，命名为决策函数$F_{N_Q}(\bullet|\theta_Q):R^{H_{net}\times W_{net}\times 3}\rightarrow R^7$，其中·表示网络的输入图像，实际计算中会输入不同的图像；R表示实数，函数意义将维度为$H_{net}\times W_{net}\times 3$的实数空间图像映射到7维实数空间的向量输出；7维向量输出对应摄像机7种操作的决策值，7种操作分别为："上转"、"下转"、"左转"、"右转"、"放大"、"缩小"和"不变"；

(2-2)设置一个深度神经网络，其网络结构与决策网络N_Q结构完全相同，命名为靶标网络N_T，网络的参数表示为θ_T，令$\theta_T=\theta_Q$，对应的靶标函数为$F_{N_T}(\bullet|\theta_T):R^{H_{net}\times W_{net}\times 3}\rightarrow R^7$，其中·表示网络的输入图像，实际计算中会输入不同的图像；R表示实数，函数意义为将维度为$H_{net}\times W_{net}\times 3$的实数空间图像映射到7维实数空间的向量输出；

(2-3)设置一个存储M_{buffer}组数据的缓存区B，设置缓存区B中当前存储样本编号$M_{sample}=0$；

(2-4)设置一个训练计数器$c_1=0$；

(2-5)设置一个连续成功定位计数器$c_2=0$；

(2-5)设置当前时间$t=0$；

(2-6)初始化摄像机到常规位置，视野最大化，开始一次定位试验；

(2-7)摄像机采集当前时刻的图像，采用双线性插值法，把图像大小变换为$H_{net}*W_{net}$，得到t时刻的RGB图像I_t，将图像I_t输入评价网络N_R，得到图像I_t的评价值$s_t=F_{N_R}(I_t|\theta_R)$；

(2-8)从"上转"、"下转"、"左转"、"右转"、"放大"、"缩小"和"不变"七种操作中根据以下法则挑选一种操作，记为a_t：

(2-9)摄像机执行步骤(2-8)选择的操作a_t，获得新的图像，采用双线性插值法，把图像大小变换为$H_{net}*W_{net}$，得到t+1时刻的RGB图像I_{t+1}；将图像I_{t+1}输入评价网络N_R，得到图像I_{t+1}的评价值$s_{t+1}=F_{N_R}(I_{t+1}|\theta_R)$；

(2-10)计算当前操作的回报值，记为r_t：

(2-10-1)如果执行操作a_t为"不变"，根据图像I_{t+1}的评价值s_{t+1}计算回报值，若评价值$s_{t+1}>\beta$，则回报值$r_t=P_s$；若评价值$s_{t+1}\leqslant\beta$，则当前操作的回报值$r_t=-P_s$；其中P_s为正数；

(2-10-2)如果执行操作a_t为其余任意一种操作，则根据图像I_t和I_{t+1}的评价差值计算回报值，评价差值$\Delta s=s_{t+1}-s_t$，如果$\Delta s>0$，则回报值$r_t=P_g$，否则，回报值$r_t=-P_g$；其中P_g都为正数；

(2-11)计算当前操作的终止标志，记为u_t：如果执行操作a_t为"不变"，则$u_t=0$；否则，$u_t=1$；

(2-12)将图像I_t，执行操作a_t，回报值r_t，终止标志u_t，图像I_{t+1}组成一组五元组$(I_t,a_t,r_t,u_t,I_{t+1})$存入缓存区B中编号$M_{sample}$的存储空间更新编号$M_{sample}$，如果$M_{sample}\geqslant M_{buffer}$，则$M_{sample}=0$，否则，$M_{sample}=M_{sample}+1$；

(2-13)如果缓存区B中存储的四元组数量小于M_{start}，则跳转到步骤(2-21)；否则，转步骤(2-14)，其中M_{start}为开始训练时的样本数；

(2-14)开始决策网络N_Q训练，从缓存区B中随机选取$M_{Q,b}$组五元组样本数据，将$M_{Q,b}$组五元组数据重新标号，记为$(I^j,a^j,r^j,u^j,\hat{I}^j)$，其中$j=1,2,3...M_{Q,b}$；$M_{Q,b}$为每次训练决策网络$N_Q$

选取的样本批量数；

(2-15)采用靶标网络N_T，计算每个样本的靶标函数值$F_{N_T}(\hat{I}^j|\theta_T)$，定义靶标估计值$t^j=r^j+\gamma u^j \max F_{N_T}(\hat{I}^j|\theta_T)$，其中γ为折损参数；j＝1,2,3...$M_{Q,b}$；

(2-16)计算当前决策网络N_Q对每个样本的七个操作的决策值$F_{N_Q}(I^j|\theta_Q)$，其中j＝1,2,3...$M_{Q,b}$；

(2-17)选择每个样本的七个决策值$F_{N_Q}(I^j|\theta_Q)$中，对应操作a^j的决策值，记为$C(F_{N_Q}(I^j|\theta_Q),a^j)$；

(2-18)定义决策网络N_Q的优化目标为$J_Q=\frac{1}{M_{Q,b}}\sum_{i=1}^{M_{Q,b}}(C(F_{N_Q}(I^j|\theta_Q),a^j)-t^j)^2$，计算优化目标对评价网络参数$\theta_Q$的梯度为$\frac{\partial J_Q}{\partial \theta_Q}$；

(2-19)采用随机梯度下降法，更新决策网络参数$\theta_Q:=\theta_Q-\alpha_Q\frac{\partial J_Q}{\partial \theta_Q}$，其中$\alpha_R$为决策网络的学习率；

(2-20)更新计数值c_1＝c_1+1；如果c_1>Count，更新靶标网络N_T的参数θ_T＝θ_Q，清零c_1＝0，否则，不更新靶标网络N_T的参数；

(2-21)如果操作a_t为“不变”，评价值s_{t+1}>β，则更新连续成功定位计数器c_2＝c_2+1；如果操作a_t为“不变”，评价值s_{t+1}≤β，则清零连续成功定位计数器c_2＝0；如果操作a_t为 其余操作，则不更新c_2；

(2-22)如果操作a_t为“不变”，则此次定位试验结束，设置t＝0重新开始计时，跳转到步骤(2-23)；如果执行操作a_t为其余六种操作之一，则更新时间t＝t+1，跳转到步骤(2-7)，继续此次目标定位试验；

(2-23)判断网络训练是否完成，如果连续成功定位计数器c_2>C_Q，则完成决策网络N_Q训练，其中C_Q为设置的连续成功次数的阈值；否则，继续训练，跳转到步骤(2-6)。

4.如权利要求3所述方法其特征在于，所述步骤(1-2)具体步骤如下：

(1-2-1)摄像机对含有目标的场景进行图像采集，目标被拍摄的角度和大小随机，采集到的图像为RGB图像，高为H_{origin}，宽为W_{origin}，此值由实际相机决定，总共采集M_{origin}张图像，记为IO_i，其中i＝1,2,…,M_{origin},M_{origin}；

(1-2-2)以原始图像的左上角为坐标原点，向下为x轴，向右为y轴；使用矩形框对每一张采集到的图像中的目标位置进行标注，将矩形框表示为$((x_0,y_0);(x_1,y_2))_i$其中i＝1,2,…,M_{origin},(x_0,y_0)和(x_1,y_1)为矩形框左上角和右下角在图像中的坐标；

(1-2-3)从每一张原始图像中遍历截取高h，宽w的所有图像，其中h取值遍历区间[$H_{origin}/2$,H_{origin}]内的所有整数，w取值遍历区间[$W_{origin}/2$,W_{origin}]内的所有整数，每张原始图像可以截取M_{cut}张图像，总截取图像数为$M_{origin}\times M_{cut}$，获取的截取图像记为$IC_{i,j}$，其中i＝1,2,…,M_{origin}，j＝1,2,…,M_{cut}；

(1-2-4)计算截取图像$IC_{i,j}$和原始图像IO_i中目标所在的矩形框$((x_0,y_0);(x_1,y_2))_i$的

面积交并比作为每一张截取图像$IC_{i,j}$的品质分数$s_{i,j}$，即：如果截取图像$IC_{i,j}$的面积为$S_{IC_{i,j}}$，矩形框$((x_0,y_0);(x_1,y_2))_i$的面积为S_{IO_i}，两者相交部分的面积为$S_{M_{i,j}}$，则每一张截取图像$IC_{i,j}$的品质分数$s_{i,j}=\frac{S_{M_{i,j}}}{S_{IC_{i,j}}+S_{IO_i}-S_{M_{i,j}}}$；

(1-2-5)通过双线性插值法，将所有截取图像的大小变化为$H_{net}*W_{net}$，并对所有截取图像重新编号后获得训练数据集$D=\{d_1,d_1,\cdots,d_{M_{train}}\}$，其中$M_{train}=M_{origin}\times M_{cut}$为数据集 中图像数量，数据集D中每个样本对应的品质分数作为训练数据集的标签，记为 $S=\{s_1,s_2,\cdots,s_{M_{train}}\}$。

5.如权利要求3所述方法，其特征在于，所述步骤(2-8)具体包括以下步骤：

(2-8-1)产生一个在区间[0,1]之间随机数q，如果$q>\varepsilon$，则将图像I_t输入决策网络N_Q中，得到7维的决策网络输出决策值$F_{N_Q}(I_t|\theta_Q)$，7个决策值分别对应“上转”、“下转”、“左转”、“右转”、“放大”、“缩小”和“不变”七种操作，选择7个决策值中最大值对应的操作，作为选择的操作a_t；其中，ε为训练中采取随机策略的概率，取值0～1之间；

(2-8-2)如果(2-8-1)中产生的随机数$q\leqslant\varepsilon$，根据评价值y_t选择操作如下：如果$y_t>\beta$，选择“不变”操作作为操作a_t，如果$y_t\leqslant\beta$，从除去“不变”操作外的其余6种操作中随机选择1种操作作为操作a_t；β为设置的摄像机成功定位目标对应的评价分界值，取值0～1之间。

6.如权利要求1所述方法，其特征在于，所述步骤(3)的具体步骤如下：

(3-1)摄像机采集当前图像，采用双线性插值法，把图像大小变换为$H_{net}*W_{net}$，得到图像I；

(3-2)将图像I输入决策网络中，得到7种操作的决策值$F_{N_Q}(I|\theta_Q)$，选择7种操作中决策值最大的操作a；

(3-3)摄像机执行操作a；

(3-4)重复步骤(3-1)～步骤(3-3)，根据学习到的决策网络完成目标定位任务。

一种基于深度强化学习的摄像机主动目标定位方法

技术领域

[0001] 本发明涉及一种基于深度强化学习的摄像机主动目标定位方法，属于模式识别技术领域和摄像机主动定位技术领域。

背景技术

[0002] 近年来，摄像机已经越来越多的应用于生产生活中，如：安保监控，车辆检测，目标跟踪，人脸识别。现阶段的应用场景中，摄像机提供图像信息，使用人工监控或目标检测算法等方式实现相应应用。在整个摄像机图像采集过程中，摄像机固定不动或者按照指定的路线循环调整角度，不能根据实际场景，主动调整视野，对目标进行主动定位。

[0003] 现有的技术文献中，发明专利“摄像机红外主动跟踪装置及采用该装置的摄像机控制系统”，公开号为102376156A，通过接收红外发射装置发出的红外信号并根据该信号对目标位置进行分析判定的目标信号拾取器，调整摄像机位置。该方法的缺点是，只能应用于对红外信号能够判别的目标定位场景，不能根据不同的应用做出相应调整。另外，该方法需要添加额外的红外装置，不是直接根据图像信息做出调整。

[0004] 深度神经网络包含多层神经网络，后一层神经网络的输入是前一层神经网络的输出，一般情况下每一层神经网络都会有采用一种非线性的激励函数，也称激活函数，常用的激励函数有，ReLU函数：ReLU(a)＝max(0,a)，其中a为输入量，如果输入量a小于0则输出为0，否则输出其本身；Sigmoid函数：$\mathrm{Sigmoid}(a)=\frac{1}{1+e^{-a}}$，其中e为自然常数。

发明内容

[0005] 本发明的目的是提出一种基于深度强化学习的摄像机主动目标定位方法，提供一种在图像采集应用中摄像机主动调整进行目标定位的方法，该方法基于深度强化学习算法，提高了采集图像的质量。本发明摄像机可以通过“上转”、“下转”、“左转”、“右转”、“放大”、“缩小”和“不变”七种操作方式定位目标物体，利用深度神经网络作为强化学习值函数的逼近器，将图像信息输入深度神经网络，从而确定当前摄像机应该做出何种操作来定位目标所在位置。

[0006] 本发明提出的一种基于深度强化学习的摄像机主动目标定位方法，其特征在于，该方法包括以下步骤：

[0007] (1)训练一个评价摄像机定位效果的深度神经网络，将该网络命名为评价网络N_R由多层神经网络组成；

[0008] (2)进行多次目标定位试验，在定位实验过程中，训练一个拟合强化学习值函数的深度神经网络，将该网络命名为决策网络N_Q，通过决策网络N_Q判断摄像机7种操作的优劣；

[0009] (3)完成决策网络N_Q训练后，采用决策网络N_Q根据摄像机当前获取的图像信息，对摄像机操作做出决策。

[0010] 本发明提出的基于深度强化学习的摄像机主动目标定位方法的技术特点及有益

效果：

[0011] 为了实现摄像机对目标主动定位的应用，本发明结合了深度神经网络算法和强化学习算法，根据摄像机拍摄到的图像，控制摄像机转动，从而实现目标定位的摄像机控制系统。强化学习又称增强学习，通过不断试错积累经验，根据积累的经验优化控制策略实现完成目标的目的。将强化学习算法应用于摄像机主动定位，摄像机获取的图像作为学习的信息来源，系统需要有很好地处理图像数据的能力。深度神经网络能够有效地提取图像特征，而且可以通过学习的方式优化特征，使得特征适应于当前任务。

[0012] 本方法具有以下有益效果：

[0013] 1、本发明中的用于摄像机主动目标定位方法，决策网络根据当前图像信息，选择摄像机操作，完成目标定位，定位过程主动完成，不需要人为参与。

[0014] 2、本发明中摄像机主动定位目标，对不同的目标定位任务，只需训练不同的评价网络，其余的学习算法具有普适性，通用性。

[0015] 3、本发明采用评价网络对图像进行质量评价的方式，评价摄像机目标定位的效果，以此确定摄像机执行当前操作的回报，让摄像机从反复的试验中，自主学习实现目标定位方法。

具体实施方式

[0016] 本发明提出的基于深度强化学习的摄像机主动目标定位方法的具体实现方式，包括以下步骤：

[0017] (1)训练一个评价摄像机定位效果的深度神经网络，将该网络命名为评价网络N_R由多层神经网络组成，具体步骤如下：

[0018] (1-1)设置评价网络N_R：评价网络N_R的网络结构依次为：输入层为RGB图像，图像高为H_{net}，宽为W_{net}，(一般设置为$H_{net}=W_{net}=256$像素)，由于RGB图像为3个维度，所以输入层的维度为$H_{net}\times W_{net}\times 3$；$L_{RC}$层为卷积神经网络，激励函数为ReLU函数($L_{RC}$层数一般取值为3～7之间)；$L_{RP}$层为全连接层($L_{RP}$的层数一般取值为2～4之间)，前$L_{RP}-1$层的激励函数也为ReLU函数，最后一层全连接层激励函数为Sigmoid函数，且设置维度为1，作为评价网络N_R输出，评价网络N_R输出定义为评价值；将评价网络N_R中的所有参数统一表示为θ_R(参数为随机初始化，在训练阶段迭代更新)，评价网络N_R逐层的运算过程表示一个函数映射，命名为评价函数$F_{N_R}(\bullet|\theta_R):R^{H_{net}\times W_{net}\times 3}\rightarrow R$，其中·表示网络的输入图像，实际计算中会输入不同的图像；R表示实数，函数意义为将维度为$H_{net}\times W_{net}\times 3$的实数空间图像映射到1维实数空间的评价值；

[0019] (1-2)采集训练评价网络N_R的数据集：具体步骤如下：

[0020] (1-2-1)摄像机对含有目标的场景进行图像采集，目标被拍摄的角度和大小随机，采集到的图像为RGB图像，高为H_{origin}，宽为W_{origin}，此值由实际相机决定，总共采集M_{origin}张图像，记为IO_i，其中$i=1,2,\cdots,M_{origin}$，M_{origin}取值大于10000张较为合适；

[0021] (1-2-2)以原始图像的左上角为坐标原点，向下为x轴，向右为y轴；使用矩形框对每一张采集到的图像中的目标位置进行标注，将矩形框表示为$((x_0,y_0);(x_1,y_2))_i$其中$i=1,2,\cdots,M_{origin}$，(x_0,y_0)和(x_1,y_1)为矩形框左上角和右下角在图像中的坐标；

[0022] (1-2-3)从每一张原始图像中遍历截取高h，宽w的所有图像，其中h取值遍历区间

$[H_{origin}/2, H_{origin}]$内的所有整数，w取值遍历区间$[W_{origin}/2, W_{origin}]$内的所有整数，每张原始图像可以截取$M_{cut}$张图像，总截取图像数为$M_{origin} \times M_{cut}$，获取的截取图像记为$IC_{i,j}$，其中i＝1,2,…,$M_{origin}$，j＝1,2,…,$M_{cut}$；

[0023] (1-2-4)计算截取图像$IC_{i,j}$和原始图像IO_i中目标所在的矩形框$((x_0,y_0);(x_1,y_2))_i$的面积交并比作为每一张截取图像$IC_{i,j}$的品质分数$s_{i,j}$，即：如果截取图像$IC_{i,j}$的面积为$S_{IC_{i,j}}$，矩形框$((x_0,y_0);(x_1,y_2))_i$的面积为S_{IO_i}，两者相交部分的面积为$S_{M_{i,j}}$，则每一张截取图像$IC_{i,j}$的品质分数$s_{i,j}=\frac{S_{M_{i,j}}}{S_{IC_{i,j}}+S_{IO_i}-S_{M_{i,j}}}$；

[0024] (1-2-5)通过双线性插值法，将所有截取图像的大小变化为$H_{net}*W_{net}$，并对所有截取图像重新编号后获得训练数据集$D=\{d_1,d_1,\cdots,d_{M_{train}}\}$，其中$M_{train}=M_{origin}\times M_{cut}$为数据集中图像数量，数据集D中每个样本对应的品质分数作为训练数据集的标签，记为$S=\{s_1,s_2,\cdots,s_{M_{train}}\}$；

[0025] (1-3)从数据集D中随机挑选$M_{R,b}$组样本，记为批量样本$D_{batch}=\{d^1,d^2,\cdots,d^{M_{batch}}\}$，以及标签S中与其对应的批量标签$S_{batch}=\{s^1,s^2,\cdots,s^{M_{R,b}}\}$；一般批量大小$M_{R,b}$取值为100；

[0026] (1-4)根据步骤(1-1)的评价函数计算评价网络N_R对批量样本D_{batch}的评价值$F_{N_R}(d^i|\theta_R)$，其中i＝1,2,…,$M_{R,b}$；

[0027] (1-5)定义评价网络N_R的优化目标为$J_R=\frac{1}{M_{R,b}}\sum_{i=1}^{M_{R,b}}(F_{N_R}(d^i|\theta_R)-s^i)^2$，其中$F_{N_R}(d^i|\theta_R)$为第i个样本$d^i$输入网络后输出的评价值，将最小化评价值和标签之间的欧式距离作为优化目标训练网络，计算优化目标对评价网络参数θ_R的梯度为$\frac{\partial J_R}{\partial \theta_R}$；

[0028] (1-6)采用随机梯度下降法，更新评价网络参数$\theta_R:=\theta_R-\alpha_R\frac{\partial J_R}{\partial \theta_R}$，其中$\alpha_R$为评价网络的学习率，一般设置为0.01；

[0029] (1-7)重复上述步骤(1-3)～(1-6)，不断更新评价网络参数θ_R，直到评价网络收敛，完成评价网络训练，评价网络收敛的依据是连续C_R次优化目标J小于阈值η_R，一般C_R取100次，阈值η_R取0.05；

[0030] (2)进行多次目标定位试验，在定位实验过程中，训练一个拟合强化学习值函数的深度神经网络，将该网络命名为决策网络N_Q，通过决策网络N_Q判断摄像机7种操作的优劣，具体步骤如下：

[0031] (2-1)设置决策网络N_Q结构依次为：输入层为RGB图像，图像高为H_{net}，宽为W_{net}，与评价网络相同；L_{QC}层为卷积神经网络，激励函数为ReLU函数(L_{QC}一般取值为3～7之间)；L_{QP}层为全连接层(L_{QP}一般取值为2～4之间)，前L_{QP}-1层的激励函数也为ReLU函数，最后一层全连接层无激励函数，设置维度为7，作为网络输出，将决策网络N_Q中的全部参数统一表示为θ_Q(参数为随机初始化，在训练阶段迭代更新)，决策网络N_Q逐层的运算过程表示为一个函数映射，命名为决策函数$F_{N_Q}(\bullet|\theta_Q):R^{H_{net}\times W_{net}\times 3}\rightarrow R^7$，其中•表示网络的输入图像，实际计算中

会输入不同的图像;R表示实数,函数意义将维度为$H_{net}\times W_{net}\times 3$的实数空间图像映射到7维实数空间的向量输出;7维向量输出对应摄像机7种操作的决策值,7种操作分别为:“上转”、“下转”、“左转”、“右转”、“放大”、“缩小”和“不变”;

[0032] (2-2)设置一个深度神经网络,其网络结构与决策网络N_Q结构完全相同,命名为靶标网络N_T,网络的参数表示为θ_T,令$\theta_T=\theta_Q$,对应的靶标函数为$F_{N_T}(\bullet|\theta_T):R^{H_{net}\times W_{net}\times 3}\rightarrow R^7$,其中·表示网络的输入图像,实际计算中会输入不同的图像;R表示实数,函数意义为将维度为$H_{net}\times W_{net}\times 3$的实数空间图像映射到7维实数空间的向量输出;

[0033] (2-3)设置一个可以存储M_{buffer}组数据的缓存区B,设置缓存区B中当前存储样本编号$M_{sample}=0$;

[0034] (2-4)设置一个训练计数器$c_1=0$;

[0035] (2-5)设置一个连续成功定位计数器$c_2=0$;

[0036] (2-5)设置当前时间$t=0$;

[0037] (2-6)初始化摄像机到常规位置,视野最大化,开始一次定位试验;

[0038] (2-7)摄像机采集当前时刻的图像,采用双线性插值法,把图像大小变换为$H_{net}*W_{net}$,得到t时刻的RGB图像I_t,将图像I_t输入评价网络N_R,得到图像I_t的评价值$s_t=F_{N_R}(I_t|\theta_R)$;

[0039] (2-8)从“上转”、“下转”、“左转”、“右转”、“放大”、“缩小”和“不变”七种操作中根据以下法则挑选一种操作,记为a_t:

[0040] (2-8-1)产生一个在区间[0,1]之间随机数q,如果$q>\varepsilon$,则将图像I_t输入决策网络N_Q中,得到7维的决策网络输出决策值$F_{N_Q}(I_t|\theta_Q)$,7个决策值分别对应“上转”、“下转”、“左转”、“右转”、“放大”、“缩小”和“不变”七种操作,选择7个决策值中最大值对应的操作,作为选择的操作a_t;其中,ε为训练中采取随机策略的概率,取值0～1之间,一般取值为0.1;

[0041] (2-8-2)如果(2-8-1)中产生的随机数$q\leq\varepsilon$,根据评价值y_t选择操作如下:如果$y_t>\beta$,选择“不变”操作作为操作a_t,如果$y_t\leq\beta$,从除去“不变”操作外的其余6种操作中随机选择1种操作作为操作a_t;β为设置的摄像机成功定位目标对应的评价分界值,取值0～1之间,一般取值为0.6;

[0042] (2-9)摄像机执行步骤(2-8)选择的操作a_t,获得新的图像,采用双线性插值法,把图像大小变换为$H_{net}*W_{net}$,得到t+1时刻的RGB图像I_{t+1};将图像I_{t+1}输入评价网络N_R,得到图像I_{t+1}的评价值$s_{t+1}=F_{N_R}(I_{t+1}|\theta_R)$;

[0043] (2-10)计算当前操作的回报值,记为r_t:

[0044] (2-10-1)如果执行操作a_t为“不变”,根据图像I_{t+1}的评价值s_{t+1}计算回报值,若评价值$s_{t+1}>\beta$,则回报值$r_t=P_s$;若评价值$s_{t+1}\leq\beta$,则当前操作的回报值$r_t=-P_s$;其中P_s为正数,一般P_s取值为1;

[0045] (2-10-2)如果执行操作a_t为其余任意一种操作,则根据图像I_t和I_{t+1}的评价差值计算回报值,评价差值$\Delta s=s_{t+1}-s_t$,如果$\Delta s>0$,则回报值$r_t=P_g$,否则,回报值$r_t=-P_g$;其中P_g都为正数,一般P_g取值为0.1;

[0046] (2-11)计算当前操作的终止标志,记为u_t:如果执行操作a_t为“不变”,则$u_t=0$;否则,$u_t=1$;

[0047] (2-12)将图像I_t,执行操作a_t,回报值r_t,终止标志u_t,图像I_{t+1}组成一组五元组(I_t,a_t,r_t,u_t,I_{t+1})存入缓存区B中编号M_{sample}的存储空间更新编号M_{sample},如果$M_{sample}\geq M_{buffer}$,则$M_{sample}=0$,否则,$M_{sample}=M_{sample}+1$;

[0048] (2-13)如果缓存区B中存储的四元组数量小于M_{start},则跳转到步骤(2-21);否则,转步骤(2-14),其中M_{start}为开始训练时的样本数,一般取值为1000;

[0049] (2-14)开始决策网络N_Q训练,从缓存区B中随机选取$M_{Q,b}$组五元组样本数据,将$M_{Q,b}$组五元组数据重新标号,记为$(I^j,a^j,r^j,u^j,\hat{I}^j)$,其中j=1,2,3...$M_{Q,b}$;$M_{Q,b}$为每次训练决策网络$N_Q$选取的样本批量数,一般取值为32;

[0050] (2-15)采用靶标网络N_T,计算每个样本的靶标函数值$F_{N_T}(\hat{I}^j|\theta_T)$,定义靶标估计值$t^j=r^j+\gamma u^j\max F_{N_T}(\hat{I}^j|\theta_T)$,其中γ为折损参数,一般设置为0.99;j=1,2,3...$M_{Q,b}$;

[0051] (2-16)计算当前决策网络N_Q对每个样本的七个操作的决策值$F_{N_Q}(I^j|\theta_Q)$,其中j=1,2,3...$M_{Q,b}$;

[0052] (2-17)选择每个样本的七个决策值$F_{N_Q}(I^j|\theta_Q)$中,对应操作a^j的决策值,记为$C(F_{N_Q}(I^j|\theta_Q),a^j)$;

[0053] (2-18)定义决策网络N_Q的优化目标为$J_Q=\frac{1}{M_{Q,b}}\sum_{i=1}^{M_{Q,b}}(C(F_{N_Q}(I^j|\theta_Q),a^j)-t^j)^2$,计算优化目标对评价网络参数$\theta_Q$的梯度为$\frac{\partial J_Q}{\partial\theta_Q}$;

[0054] (2-19)采用随机梯度下降法,更新决策网络参数$\theta_Q:=\theta_Q-\alpha_Q\frac{\partial J_Q}{\partial\theta_Q}$,其中$\alpha_R$为决策网络的学习率,一般设置为0.0001;

[0055] (2-20)更新计数值$c_1=c_1+1$;如果$c_1>Count$,更新靶标网络N_T的参数$\theta_T=\theta_Q$,清零$c_1=0$,否则,不更新靶标网络N_T的参数;

[0056] (2-21)如果操作a_t为“不变”,评价值$s_{t+1}>\beta$,则更新连续成功定位计数器$c_2=c_2+1$;如果操作a_t为“不变”,评价值$s_{t+1}\leq\beta$,则清零连续成功定位计数器$c_2=0$;如果操作a_t为其余操作,则不更新c_2;

[0057] (2-22)如果操作a_t为“不变”,则此次定位试验结束,设置t=0重新开始计时,跳转到步骤(2-23);如果执行操作a_t为其余六种操作之一,则更新时间t=t+1,跳转到步骤(2-7),继续此次目标定位试验;

[0058] (2-23)判断网络训练是否完成,如果连续成功定位计数器$c_2>C_Q$,则完成决策网络N_Q训练,其中C_Q为设置的连续成功次数的阈值,一般取值为100次;否则,继续训练,跳转到步骤(2-6);

[0059] (3)完成决策网络N_Q训练后,采用决策网络N_Q根据摄像机当前获取的图像信息,对摄像机操作做出决策;具体步骤如下:

[0060] (3-1)摄像机采集当前图像,采用双线性插值法,把图像大小变换为$H_{net}*W_{net}$,得到图像I;

[0061] (3-2)将图像I输入决策网络中,得到7种操作的决策值$F_{N_Q}(I|\theta_Q)$,选择7种操作中决策值最大的操作a;

[0062] (3-3)摄像机执行操作a;

[0063] (3-4)重复步骤(3-1)～步骤(3-3),根据学习到的决策网络完成目标定位任务。