(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 106094516 A
(43)申请公布日 2016.11.09

(21)申请号 201610402319.6

(22)申请日 2016.06.08

(71)申请人 南京大学
地址 210093 江苏省南京市鼓楼区汉口路22号

(72)发明人 陈春林 侯跃南 刘力锋 魏青 徐旭东 朱张青 辛博 马海兰

(74)专利代理机构 南京天翼专利代理有限责任公司 32112
代理人 于忠洲

(51)Int.Cl.
G05B 13/04(2006.01)

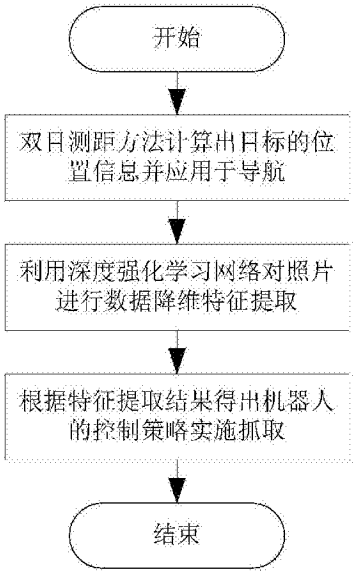
权利要求书2页 说明书7页 附图3页

(54)发明名称

一种基于深度强化学习的机器人自适应抓取方法

(57)摘要

本发明提供了一种基于深度强化学习的机器人自适应抓取方法，步骤包括：在距离待抓取目标一定距离时，机器人通过前部的摄像头获取目标的照片，再根据照片利用双目测距方法计算出目标的位置信息，并将计算出的位置信息用于机器人导航；当目标进入机械手臂抓范围内时，再通过前部的摄像头拍摄目标的照片，并利用预先训练过的基于DDPG的深度强化学习网络对照片进行数据降维特征提取；根据特征提取结果得出机器人的控制策略，机器人利用控制策略来控制运动路径和机械手臂的位姿，从而实现目标的自适应抓取。该抓取方法能够对大小形状不同、位置不固定的物体实现自适应抓取，具有良好的市场应用前景。



CN 106094516 A

1.一种基于深度强化学习的机器人自适应抓取方法，其特征在于，包括如下步骤：

步骤1，在距离待抓取目标一定距离时，机器人通过前部的摄像头获取目标的照片，再根据照片利用双目测距方法计算出目标的位置信息，并将计算出的位置信息用于机器人导航；

步骤2，机器人根据导航进行移动，当目标进入机械手臂抓范围内时，再通过前部的摄像头拍摄目标的照片，并利用预先训练过的基于DDPG的深度强化学习网络对照片进行数据降维特征提取；

步骤3，根据特征提取结果得出机器人的控制策略，机器人利用控制策略来控制运动路径和机械手臂的位姿，从而实现目标的自适应抓取。

2.根据权利要求1所述的基于深度强化学习的机器人自适应抓取方法，其特征在于，步骤1中根据照片利用双目测距方法计算出目标的位置信息的具体步骤为：

步骤1.1，获取摄像头的焦距f、左右两个摄像头的中心距T_x以及目标点在左右两个摄像头的像平面的投影点到各自像平面最左侧的物理距离x^l和x^r，左右两个摄像头对应的左侧的像平面和右侧的像平面均为矩形平面，且位于同一成像平面上，左右两个摄像头的光心投影分别位于相应像平面的中心处，则视差d为：

$$d=x^l-x^r \quad (1)$$

步骤1.2，利用三角形相似原理建立Q矩阵为：

$$Q=\begin{bmatrix}1 & 0 & 0 & -c_x\\ 0 & 1 & 0 & -c_y\\ 0 & 0 & 0 & f\\ 0 & 0 & -\frac{1}{T_x} & \frac{c_x-c_x'}{T_x}\end{bmatrix} \quad (2)$$

$$Q\begin{bmatrix}x\\ y\\ d\\ 1\end{bmatrix}=\begin{bmatrix}x-c_x\\ y-c_y\\ f\\ \frac{-d+c_x-c_x'}{T_x}\end{bmatrix}=\begin{bmatrix}X\\ Y\\ Z\\ W\end{bmatrix} \quad (3)$$

式(2)和(3)中，(X,Y,Z)为目标点在以左摄像头光心为原点的立体坐标系中的坐标，W为旋转平移变换比例系数，(x,y)为目标点在左侧的像平面中的坐标，c_x和c_y分别为左侧的像平面和右侧的像平面的坐标系与立体坐标系中原点的偏移量，c_x'为c_x的修正值；

步骤1.3，计算得到目标点到成像平面的空间距离为：

$$Z=\frac{-T_x f}{d-(c_x-c_x')} \quad (4)$$

将左摄像头的光心所在位置作为机器人所在位置，将目标点的坐标位置信息(X,Y,Z)作为导航目的地进行机器人导航。

3.根据权利要求1或2所述的基于深度强化学习的机器人自适应抓取方法，其特征在于，步骤2中利用预先训练过的基于DDPG的深度强化学习网络对照片进行数据降维特征提取的具体步骤为：

步骤2.1,利用目标抓取过程符合强化学习且满足马尔科夫性质的条件,计算t时刻之前的观察量和动作的集合为:

$s_t=(x_1,a_1,\ldots,a_{t-1},x_t)=x_t$ (5)

式(5)中,x_t和a_t分别为t时刻的观察量以及所采取的动作;

步骤2.2,利用策略值函数来描述抓取过程的预期收益为:

$Q^{\pi}(s_t,a_t)=E[R_t|s_t,a_t]$ (6)

式(6)中,$R_t=\sum_{i=t}^{T}\gamma^{i-t}r(s_i,a_i)$为时刻t获得的打过折扣以后的未来收益总和,$\gamma\in[0,1]$为折扣因子,$r(s_t,a_t)$为时刻t的收益函数,T为抓取结束的时刻,π为抓取策略;

由于抓取的目标策略π是预设确定的,记为函数μ:S←A,S为状态空间,A为N维度的动作空间,同时利用贝尔曼方程处理式(6)有:

$$Q^{\mu}(s_t,a_t)=E_{s_{t+1}\sim E}[r(s_t,a_t)+\gamma Q^{\mu}(s_{t+1},\mu(s_{t+1}))] \quad (7)$$

式(7)中,$s_{t+1}\sim E$表示t+1时刻的观察量是从环境E中获得的,$\mu(s_{t+1})$表示t+1时刻从观察量通过函数μ所映射到的动作;

步骤2.3,利用最大似然估计的原则,通过最小化损失函数来更新网络权重参数为θ^Q的策略评估网络$Q(s,a|\theta^Q)$,所采用的损失函数为:

$L(\theta^Q)=E_{\mu'}[(Q(s_t,a_t|\theta^Q)-y_t)^2]$ (8)

式(8)中,$y_t=r(s_t,a_t)+\gamma Q(s_{t+1},\mu(s_{t+1})|\theta^Q)$为目标策略评估网络,μ'为目标策略;

步骤2.4,对于实际的参数为θ^{μ}的策略函数$\mu(s|\theta^{\mu})$,利用链式法得到的梯度为:

$$\nabla_{\theta^{\mu}}\mu\approx E_{\mu'}[\nabla_{\theta^{\mu}}Q(s,a|\theta^Q)|_{s=s_t,a=\mu(s_t|\theta^{\mu})}]$$
$$=E_{\mu'}[\nabla_a Q(s,a|\theta^Q)|_{s=s_t,a=\mu(s_t)}\nabla_{\theta^{\mu}}\mu(s|\theta^{\mu})|_{s=s_t}] \quad (9)$$

由式(9)计算得到的梯度即为策略梯度,再利用策略梯度来更新策略函数$\mu(s|\theta^{\mu})$;

步骤2.5,利用离策略算法来训练网络,网络训练中用到的样本数据从同一个样本缓冲区中得到,以最小化样本之间的关联性,同时用一个目标Q值网络来训练神经网络,即采用经验回放机制和目标Q值网络方法对于目标网络的更新,所采用的缓慢更新策略为:

$\theta^{Q'}\leftarrow\tau\theta^Q+(1-\tau)\theta^{Q'}$ (10)

$\theta^{\mu'}\leftarrow\tau\theta^{\mu}+(1-\tau)\theta^{\mu'}$ (11)

式(10)和(11)中,τ为更新率,$\tau<<1$,由此便构建了一个基于DDPG的深度强化学习网络,且为收敛的神经网络;

步骤2.6,利用构建好的深度强化学习网络对照片进行数据降维特征提取,获得机器人的控制策略。

4.根据权利要求3所述的基于深度强化学习的机器人自适应抓取方法,其特征在于,步骤2.6中的深度强化学习网络由一个图像输入层、两个卷积层、两个全连接层以及一个输出层构成,图像输入层用于输入包含待抓取物体的图像;卷积层用于提取特征,即一个图像的深层表现形式;全连接层和输出层用于构成一个深层网络,通过训练以后,输入特征信息到该深层网络即可输出控制指令,即控制机器人的机械手臂舵机角度和控制搭载小车的直流电机转速。

一种基于深度强化学习的机器人自适应抓取方法

技术领域

[0001] 本发明涉及一种机器人抓取物体的方法，尤其是一种基于深度强化学习的机器人自适应抓取方法。

背景技术

[0002] 自主机器人是高度智能化的服务型机器人，具有对外界环境的学习功能。为了实现各种基本活动(如定位、移动、抓取)的功能，需要机器人配有机械手臂和机械手爪并融合多传感器的信息来进行机器学习(如深度学习和强化学习)，与外界环境进行交互，实现其感知、决策和行动等各项功能。现在绝大多数抓取型机器人工作在待抓取物件大小、形状和位置相对固定的情况，并且抓取技术主要是基于超声波、红外和激光测距等传感器，因此使用范围很受限制，无法适应抓取环境更为复杂、抓取物件大小、形状和位置不固定的情况；目前，现有的视觉型机器人技术很难解决输入的视觉信息维度高、数据量大的“维数灾难”问题；并且，利用机器学习训练的神经网络也很难收敛，无法直接处理输入的图像信息。总体来说，现在的视觉型抓取服务机器人的控制技术尚未达到令人满意的结果，尤其在实用中还需要进一步优化。

发明内容

[0003] 本发明要解决的技术问题是现有的无法适应抓取环境更为复杂、抓取物件大小、形状和位置不固定的情况。

[0004] 为了解决上述技术问题，本发明提供了一种基于深度强化学习的机器人自适应抓取方法，包括如下步骤：

[0005] 步骤1，在距离待抓取目标一定距离时，机器人通过前部的摄像头获取目标的照片，再根据照片利用双目测距方法计算出目标的位置信息，并将计算出的位置信息用于机器人导航；

[0006] 步骤2，机器人根据导航进行移动，当目标进入机械手臂抓范围内时，再通过前部的摄像头拍摄目标的照片，并利用预先训练过的基于DDPG的深度强化学习网络对照片进行数据降维特征提取；

[0007] 步骤3，根据特征提取结果得出机器人的控制策略，机器人利用控制策略来控制运动路径和机械手臂的位姿，从而实现目标的自适应抓取。

[0008] 作为本发明的进一步限定方案，步骤1中根据照片利用双目测距方法计算出目标的位置信息的具体步骤为：

[0009] 步骤1.1，获取摄像头的焦距f、左右两个摄像头的中心距T_x以及目标点在左右两个摄像头的像平面的投影点到各自像平面最左侧的物理距离x^l和x^r，左右两个摄像头对应的左侧的像平面和右侧的像平面均为矩形平面，且位于同一成像平面上，左右两个摄像头的光心投影分别位于相应像平面的中心处，则视差d为：

[0010] $$d=x^l-x^r \quad (1)$$

[0011] 步骤1.2,利用三角形相似原理建立Q矩阵为:

[0012] $$Q=\begin{bmatrix}1 & 0 & 0 & -c_x\\ 0 & 1 & 0 & -c_y\\ 0 & 0 & 0 & f\\ 0 & 0 & -\frac{1}{T_x} & \frac{c_x-c_x'}{T_x}\end{bmatrix} \quad (2)$$

[0013] $$Q\begin{bmatrix}x\\ y\\ d\\ 1\end{bmatrix}=\begin{bmatrix}x-c_x\\ y-c_y\\ f\\ \frac{-d+c_x-c_x'}{T_x}\end{bmatrix}=\begin{bmatrix}X\\ Y\\ Z\\ W\end{bmatrix} \quad (3)$$

[0014] 式(2)和(3)中,(X,Y,Z)为目标点在以左摄像头光心为原点的立体坐标系中的坐标,W为旋转平移变换比例系数,(x,y)为目标点在左侧的像平面中的坐标,c_x和c_y分别为左侧的像平面和右侧的像平面的坐标系与立体坐标系中原点的偏移量,c_x'为c_x的修正值;

[0015] 步骤1.3,计算得到目标点到成像平面的空间距离为:

[0016] $$Z=\frac{-T_x f}{d-(c_x-c_x')} \quad (4)$$

[0017] 将左摄像头的光心所在位置作为机器人所在位置,将目标点的坐标位置信息(X,Y,Z)作为导航目的地进行机器人导航。

[0018] 作为本发明的进一步限定方案,步骤2中利用预先训练过的基于DDPG的深度强化学习网络对照片进行数据降维特征提取的具体步骤为:

[0019] 步骤2.1,利用目标抓取过程符合强化学习且满足马尔科夫性质的条件,计算t时刻之前的观察量和动作的集合为:

[0020] $s_t=(x_1,a_1,\ldots,a_{t-1},x_t)=x_t$ (5)

[0021] 式(5)中,x_t和a_t分别为t时刻的观察量以及所采取的动作;

[0022] 步骤2.2,利用策略值函数来描述抓取过程的预期收益为:

[0023] $Q^{\pi}(s_t,a_t)=E[R_t|s_t,a_t]$ (6)

[0024] 式(6)中,$R_t=\sum_{i=t}^{T}\gamma^{i-t}r(s_i,a_i)$为时刻t获得的打过折扣以后的未来收益总和,$\gamma\in[0,1]$为折扣因子,$r(s_t,a_t)$为时刻t的收益函数,T为抓取结束的时刻,π为抓取策略;

[0025] 由于抓取的目标策略π是预设确定的,记为函数μ:S←A,S为状态空间,A为N维度的动作空间,同时利用贝尔曼方程处理式(6)有:

[0026] $$Q^{\mu}(s_t,a_t)=E_{s_{t+1}\sim E}[r(s_t,a_t)+\gamma Q^{\mu}(s_{t+1},\mu(s_{t+1}))] \quad (7)$$

[0027] 式(7)中,$s_{t+1}\sim E$表示t+1时刻的观察量是从环境E中获得的,$\mu(s_{t+1})$表示t+1时刻从观察量通过函数μ所映射到的动作;

[0028] 步骤2.3,利用最大似然估计的原则,通过最小化损失函数来更新网络权重参数为θ^Q的策略评估网络$Q(s,a|\theta^Q)$,所采用的损失函数为:

[0029] $L(\theta^Q)=E_{\mu'}[(Q(s_t,a_t|\theta^Q)-y_t)^2]$ (8)

[0030] 式(8)中,$y_t=r(s_t,a_t)+\gamma Q(s_{t+1},\mu(s_{t+1})|\theta^Q)$为目标策略评估网络,μ'为目标策

略；

[0031] 步骤2.4，对于实际的参数为θ^{μ}的策略函数$\mu(s|\theta^{\mu})$，利用链式法得到的梯度为：

[0032]
$$\nabla_{\theta^{\mu}}\mu \approx E_{\mu'}[\nabla_{\theta^{\mu}}Q(s,a|\theta^{Q})|_{s=s_t,a=\mu(s_t|\theta^{\mu})}]$$
$$= E_{\mu'}[\nabla_{a}Q(s,a|\theta^{Q})|_{s=s_t,a=\mu(s_t)}\nabla_{\theta^{\mu}}\mu(s|\theta^{\mu})|_{s=s_t}] \quad (9)$$

[0033] 由式(9)计算得到的梯度即为策略梯度，再利用策略梯度来更新策略函数$\mu(s|\theta^{\mu})$；

[0034] 步骤2.5，利用离策略算法来训练网络，网络训练中用到的样本数据从同一个样本缓冲区中得到，以最小化样本之间的关联性，同时用一个目标Q值网络来训练神经网络，即采用经验回放机制和目标Q值网络方法对于目标网络的更新，所采用的缓慢更新策略为：

[0035] $\theta^{Q'}\leftarrow\tau\theta^{Q}+(1-\tau)\theta^{Q'}$ (10)

[0036] $\theta^{\mu'}\leftarrow\tau\theta^{\mu}+(1-\tau)\theta^{\mu'}$ (11)

[0037] 式(10)和(11)中，τ为更新率，$\tau<<1$，由此便构建了一个基于DDPG的深度强化学习网络，且为收敛的神经网络；

[0038] 步骤2.6，利用构建好的深度强化学习网络对照片进行数据降维特征提取，获得机器人的控制策略。

[0039] 作为本发明的进一步限定方案，步骤2.6中的深度强化学习网络由一个图像输入层、两个卷积层、两个全连接层以及一个输出层构成，图像输入层用于输入包含待抓取物体的图像；卷积层用于提取特征，即一个图像的深层表现形式；全连接层和输出层用于构成一个深层网络，通过训练以后，输入特征信息到该深层网络即可输出控制指令，即控制机器人的机械手臂舵机角度和控制搭载小车的直流电机转速。将所选择的卷积层和全连接层的数量为两个的目的是既可以有效提取图像特征，又可以使得神经网络在训练时便于收敛。

[0040] 本发明的有益效果在于：(1)预训练神经网络时采用经验回放机制和随机采样确定输入的图像信息可以有效解决照片前后相关度较大不满足神经网络对于输入数据彼此独立要求的问题；(2)通过深度学习实现数据降维，采用目标Q值网络法来不断调整神经网络的权重矩阵，可以尽可能地保证训练的神经网络收敛；(3)已经训练好的基于DDPG的深度强化学习神经网络可以实现数据降维和物件特征提取，并直接给出机器人的运动控制策略，有效解决“维数灾难”问题。

附图说明

[0041] 图1为本发明的系统结构示意图；

[0042] 图2为本发明的方法流程图；

[0043] 图3为本发明的双目测距方法平面示意图；

[0044] 图4为本发明的双目测距技术立体示意图；

[0045] 图5为本发明的基于DDPG的深度强化学习网络的构成示意图。

具体实施方式

[0046] 如图1所示，本发明的一种基于深度强化学习方法的机器人自适应抓取的系统包

括:图像处理系统、无线通讯系统和机器人运动系统。

[0047] 其中,图像处理系统主要有安装在机器人前部的摄像头和matlab软件构成;无线通讯系统主要由WIFI模块构成;机器人运动系统主要由底座小车和机械手臂构成;首先需要借助动力学仿真平台预训练基于DDPG(深度确定性策略梯度)的深度强化学习网络,在此过程中通常采用经验回放机制和目标Q值网络这两种方法来确保基于DDPG的深度强化学习网络在预训练过程中能收敛,接着图像处理系统获取目标物体的图像,通过无线通讯系统将图像信息传给电脑,在机器人距离待抓取物体较远时,采用双目测距技术,以得到目标物体的位置信息并将其用于机器人的导航。

[0048] 当机器人移动至机械手臂可以抓到物体时,此时再拍摄物体照片并利用已经训练好的基于DDPG的深度强化学习网络实现数据降维提取特征并给出机器人的控制策略,最后将控制策略通过无线通讯系统传送给机器人运动系统来控制机器人的运动状态,实现目标物体的准确抓取。

[0049] 预训练时首先利用matlab软件将目标物体的RGB图像转化为灰度图像,再采用经验回放机制,使得照片前后相关度尽可能小以满足神经网络对于输入数据彼此独立的要求,最后通过随机采样来获得输入神经网络的图像;通过深度学习实现数据降维,采用目标Q值网络法来不断调整神经网络的权重矩阵,最终得到收敛的神经网络。

[0050] 机器人的控制用Arduino板实现,板上自带了WIFI模块,机械手臂由4个舵机构成,共实现4个自由度,底座小车由直流电机驱动;图像处理系统主要由摄像头及其图像传输软件和matlab为主;摄像头拍摄到的目标物体的照片将由Arduino板上的WIFI模块传输到电脑,并交由matlab处理。

[0051] 系统在工作时,步骤如下:

[0052] 步骤1,首先需要借助动力学仿真平台预训练基于DDPG(深度确定性策略梯度)的深度强化学习网络,在此过程中通常采用经验回放机制和目标Q值网络这两种方法来确保基于DDPG的深度强化学习网络在预训练过程中能收敛;

[0053] 步骤2,用安装在机器人前部的摄像头获取目标物体的图像,利用WIFI模块将图像信息传给电脑;

[0054] 步骤3,在机器人距离待抓取物体较远时,采用双目测距技术,以得到目标物体的位置信息并将其用于机器人的导航;

[0055] 步骤4,当机器人移动至机械手臂可以抓到物体时,此时再拍摄物体照片并利用已经训练好的基于DDPG的深度强化学习网络实现数据降维提取特征并给出机器人的控制策略;

[0056] 步骤5,利用WIFI模块将控制信息传送给机器人运动系统,实现目标物体的准确抓取;

[0057] 如图3和图4所示,双目测距技术主要利用了目标点在左右两幅视图上成像的横向坐标直接存在的差异(即视差)与目标点到成像平面的距离存在着反比例的关系。一般情况下,焦距的量纲是像素点,摄像头中心距的量纲由定标板棋盘格的实际尺寸和我们的输入值确定,一般是以毫米为单位(为了提高精度我们设置为0.1毫米量级),视差的量纲也是像素点。因此分子分母约去,目标点到成像平面的距离的量纲与摄像头中心距的相同。

[0058] 如图5所示,基于DDPG的深度强化学习网络主要由一个图像输入层、两个卷积层、

两个全连接层、一个输出层构成。深度网络架构用于实现数据降维，卷积层用于提取特征，输出层输出控制信息。

[0059] 如图2所示，本发明提供了一种基于深度强化学习的机器人自适应抓取方法，包括如下步骤：

[0060] 步骤1，在距离待抓取目标一定距离时，机器人通过前部的摄像头获取目标的照片，再根据照片利用双目测距方法计算出目标的位置信息，并将计算出的位置信息用于机器人导航；

[0061] 步骤2，机器人根据导航进行移动，当目标进入机械手臂抓范围内时，再通过前部的摄像头拍摄目标的照片，并利用预先训练过的基于DDPG的深度强化学习网络对照片进行数据降维特征提取；

[0062] 步骤3，根据特征提取结果得出机器人的控制策略，机器人利用控制策略来控制运动路径和机械手臂的位姿，从而实现目标的自适应抓取。

[0063] 其中，步骤1中根据照片利用双目测距方法计算出目标的位置信息的具体步骤为：

[0064] 步骤1.1，获取摄像头的焦距f、左右两个摄像头的中心距T_x以及目标点在左右两个摄像头的像平面的投影点到各自像平面最左侧的物理距离x^l和x^r，左右两个摄像头对应的左侧的像平面和右侧的像平面均为矩形平面，且位于同一成像平面上，左右两个摄像头的光心投影分别位于相应像平面的中心处，即O_l、O_r在成像平面的投影点，则视差d为：

[0065] $$d=x^l-x^r \quad (1)$$

[0066] 步骤1.2，利用三角形相似原理建立Q矩阵为：

[0067] $$Q=\begin{bmatrix}1 & 0 & 0 & -c_x\\ 0 & 1 & 0 & -c_y\\ 0 & 0 & 0 & f\\ 0 & 0 & -\frac{1}{T_x} & \frac{c_x-c_x'}{T_x}\end{bmatrix} \quad (2)$$

[0068] $$Q\begin{bmatrix}x\\ y\\ d\\ 1\end{bmatrix}=\begin{bmatrix}x-c_x\\ y-c_y\\ f\\ \frac{-d+c_x-c_x'}{T_x}\end{bmatrix}=\begin{bmatrix}X\\ Y\\ Z\\ W\end{bmatrix} \quad (3)$$

[0069] 式(2)和(3)中，(X,Y,Z)为目标点在以左摄像头光心为原点的立体坐标系中的坐标，W为旋转平移变换比例系数，(x,y)为目标点在左侧的像平面中的坐标，c_x和c_y分别为左侧的像平面和右侧的像平面的坐标系与立体坐标系中原点的偏移量，c_x'为c_x的修正值(两者数值一般相差不大，在本发明中可以认为两者近似相等)；

[0070] 步骤1.3，计算得到目标点到成像平面的空间距离为：

[0071] $$Z=\frac{-T_x f}{d-(c_x-c_x')} \quad (4)$$

[0072] 将左摄像头的光心所在位置作为机器人所在位置，将目标点的坐标位置信息(X,Y,Z)作为导航目的地进行机器人导航。

[0073] 步骤2中利用预先训练过的基于DDPG的深度强化学习网络对照片进行数据降维特

征提取的具体步骤为：

[0074] 步骤2.1，利用目标抓取过程符合强化学习且满足马尔科夫性质的条件，计算t时刻之前的观察量和动作的集合为：

[0075] $s_t=(x_1,a_1,\ldots,a_{t-1},x_t)=x_t$ (5)

[0076] 式(5)中，x_t和a_t分别为t时刻的观察量以及所采取的动作；

[0077] 步骤2.2，利用策略值函数来描述抓取过程的预期收益为：

[0078] $Q^{\pi}(s_t,a_t)=E[R_t|s_t,a_t]$ (6)

[0079] 式(6)中，$R_t=\sum_{i=t}^{T}\gamma^{i-t}r(s_i,a_i)$为时刻t获得的打过折扣以后的未来收益总和，$\gamma\in[0,1]$为折扣因子，$r(s_t,a_t)$为时刻t的收益函数，T为抓取结束的时刻，π为抓取策略；

[0080] 由于抓取的目标策略π是预设确定的，记为函数μ:S←A，S为状态空间，A为N维度的动作空间，同时利用贝尔曼方程处理式(6)有：

[0081] $$Q^{\mu}(s_t,a_t)=E_{s_{t+1}\sim E}[r(s_t,a_t)+\gamma Q^{\mu}(s_{t+1},\mu(s_{t+1}))] \quad (7)$$

[0082] 式(7)中，$s_{t+1}\sim E$表示t+1时刻的观察量是从环境E中获得的，$\mu(s_{t+1})$表示t+1

[0083] 时刻从观察量通过函数μ所映射到的动作；

[0084] 步骤2.3，利用最大似然估计的原则，通过最小化损失函数来更新网络权重参数为θ^Q的策略评估网络$Q(s,a|\theta^Q)$，所采用的损失函数为：

[0085] $L(\theta^Q)=E_{\mu'}[(Q(s_t,a_t|\theta^Q)-y_t)^2]$ (8)

[0086] 式(8)中，$y_t=r(s_t,a_t)+\gamma Q(s_{t+1},\mu(s_{t+1})|\theta^Q)$为目标策略评估网络，μ'为目标策略；

[0087] 步骤2.4，对于实际的参数为θ^{μ}的策略函数$\mu(s|\theta^{\mu})$，利用链式法得到的梯度为：

[0088] $$\nabla_{\theta^{\mu}}\mu\approx E_{\mu'}[\nabla_{\theta^{\mu}}Q(s,a|\theta^{Q})|_{s=s_t,a=\mu(s_t|\theta^{\mu})}]$$
$$=E_{\mu'}[\nabla_{a}Q(s,a|\theta^{Q})|_{s=s_t,a=\mu(s_t)}\nabla_{\theta^{\mu}}\mu(s|\theta^{\mu})|_{s=s_t}] \quad (9)$$

[0089] 由式(9)计算得到的梯度即为策略梯度，再利用策略梯度来更新策略函数$\mu(s|\theta^{\mu})$；

[0090] 步骤2.5，利用离策略算法来训练网络，网络训练中用到的样本数据从同一个样本缓冲区中得到，以最小化样本之间的关联性，同时用一个目标Q值网络来训练神经网络，即采用经验回放机制和目标Q值网络方法对于目标网络的更新，所采用的缓慢更新策略为：

[0091] $\theta^{Q'}\leftarrow\tau\theta^{Q}+(1-\tau)\theta^{Q'}$ (10)

[0092] $\theta^{\mu'}\leftarrow\tau\theta^{\mu}+(1-\tau)\theta^{\mu'}$ (11)

[0093] 式(10)和(11)中，τ为更新率，$\tau<<1$，由此便构建了一个基于DDPG的深度强化学习网络，且为收敛的神经网络；

[0094] 步骤2.6，利用构建好的深度强化学习网络对照片进行数据降维特征提取，获得机器人的控制策略；深度强化学习网络由一个图像输入层、两个卷积层、两个全连接层以及一个输出层构成，其中，所选择的卷积层和全连接层的数量为两个的目的是既可以有效提取图像特征，又可以使得神经网络在训练时便于收敛；图像输入层用于输入包含待抓取物体的图像；卷积层用于提取特征，即一个图像的深层表现形式，如一些线条、边、弧线等；全连

接层和输出层用于构成一个深层网络，通过训练以后，输入特征信息到该网络可以输出控制指令，即控制机器人的机械手臂舵机角度和控制搭载小车的直流电机转速。

[0095] 本发明预训练神经网络时采用经验回放机制和随机采样确定输入的图像信息可以有效解决照片前后相关度较大不满足神经网络对于输入数据彼此独立要求的问题；通过深度学习实现数据降维，采用目标Q值网络法来不断调整神经网络的权重矩阵，可以尽可能地保证训练的神经网络收敛；已经训练好的基于DDPG的深度强化学习神经网络可以实现数据降维和物件特征提取，并直接给出机器人的运动控制策略，有效解决“维数灾难”问题。

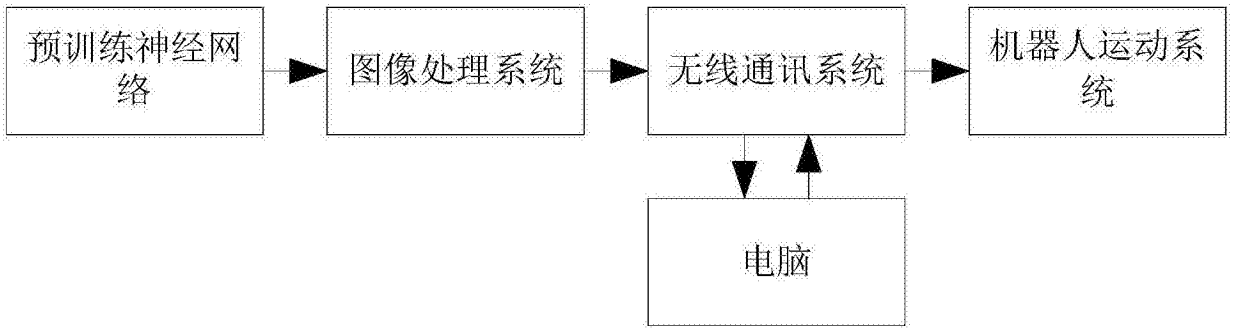

预训练神经网络
图像处理系统
无线通讯系统
机器人运动系统
电脑

图1

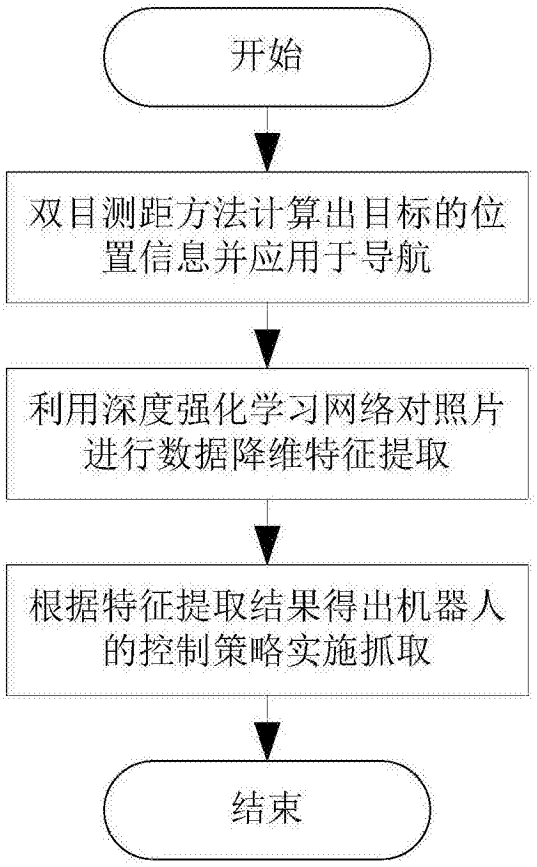

开始
双目测距方法计算出目标的位置信息并应用于导航
利用深度强化学习网络对照片进行数据降维特征提取
根据特征提取结果得出机器人的控制策略实施抓取
结束

图2

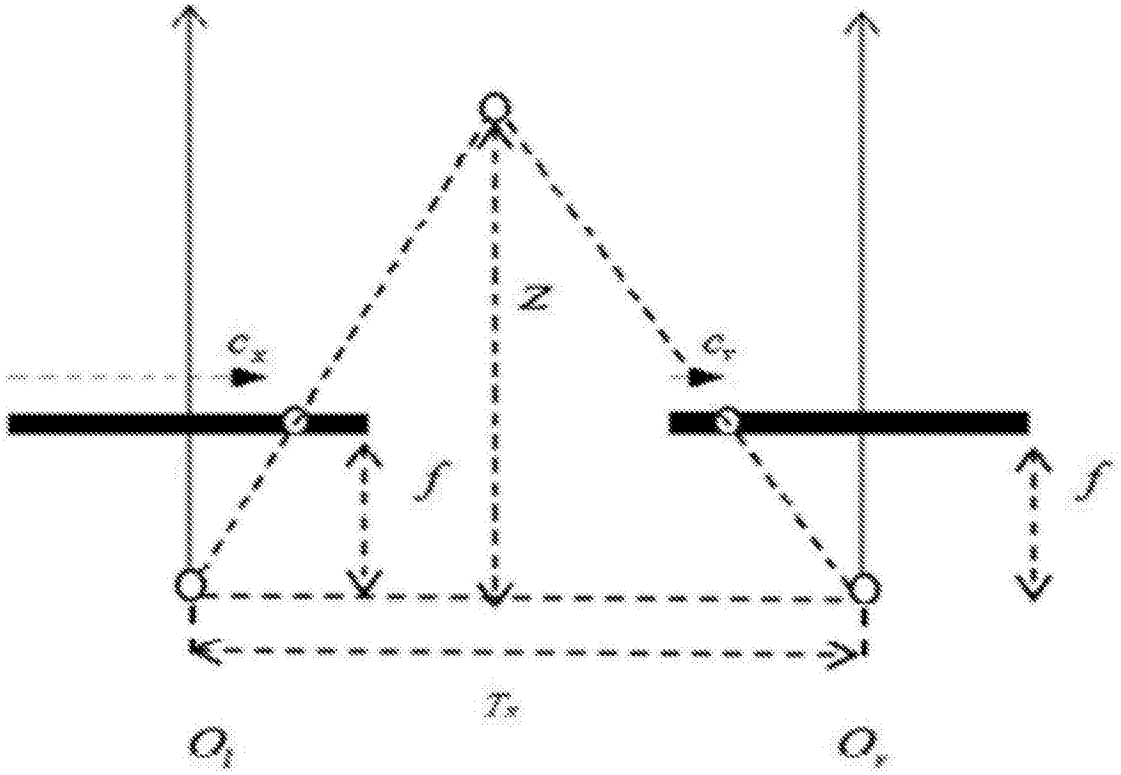


图3

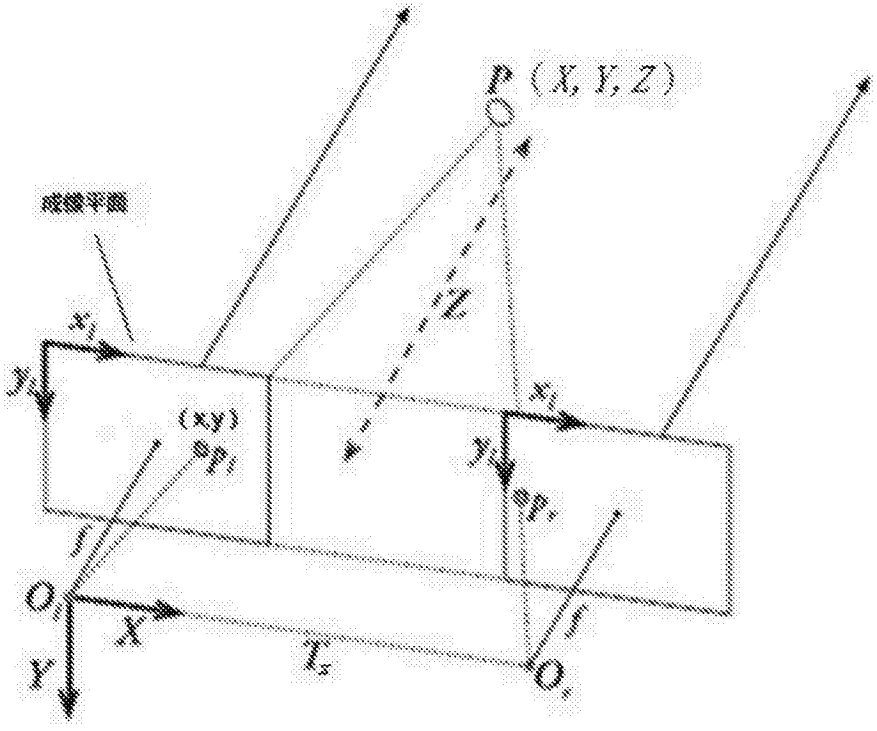


图4

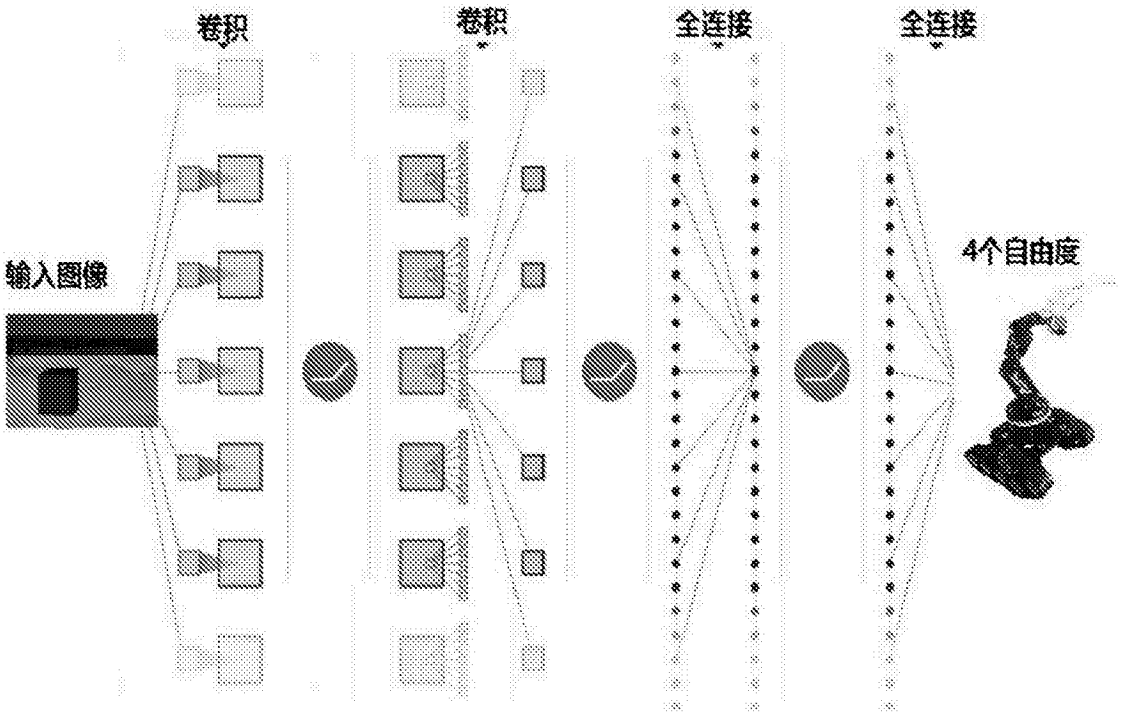


图5