(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 108393890 A
(43)申请公布日 2018.08.14

(21)申请号 201810108347.6

(22)申请日 2018.02.02

(30)优先权数据
2017-019311 2017.02.06 JP

(71)申请人 精工爱普生株式会社
地址 日本东京

(72)发明人 长谷川浩

(74)专利代理机构 北京康信知识产权代理有限责任公司 11240
代理人 张永明 玉昌峰

(51)Int.Cl.
B25J 9/16(2006.01)
B25J 19/02(2006.01)
B25J 9/00(2006.01)

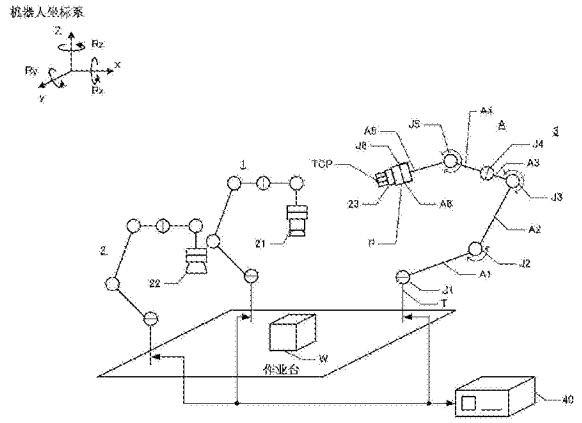
权利要求书2页 说明书37页 附图10页

(54)发明名称

控制装置、机器人以及机器人系统

(57)摘要

对充分挖掘出机器人的性能的动作参数进行设定依旧是困难的。本申请提供一种控制装置、机器人以及机器人系统。该控制装置具备:算出部,使用机械学习算出与机器人的动作相关的动作参数;以及控制部,基于算出的所述动作参数来控制所述机器人。



CN 108393890 A

1.一种控制装置,其特征在于,具备:

算出部,使用机械学习算出与机器人的动作相关的动作参数;以及

控制部,基于算出的所述动作参数控制所述机器人。

2.根据权利要求1所述的控制装置,其特征在于,

所述动作参数包含用于控制所述机器人具备的马达的伺服增益。

3.根据权利要求1或2所述的控制装置,其特征在于,

所述动作参数包含所述机器人的加减速特性。

4.根据权利要求1至3中任一项所述的控制装置,其特征在于,

所述动作参数包含用于基于所述机器人具备的惯性传感器进行控制的伺服增益。

5.根据权利要求1至4中任一项所述的控制装置,其特征在于,

所述动作参数不包含所述机器人的动作的起始点以及终点。

6.根据权利要求1至5中任一项所述的控制装置,其特征在于,

所述动作参数不包含对所述机器人示教过的示教位置。

7.根据权利要求1至6中任一项所述的控制装置,其特征在于,

所述动作参数包含所述机器人在通过阻抗控制进行动作时的阻抗参数。

8.根据权利要求1至7中任一项所述的控制装置,其特征在于,

所述算出部包含:

状态观测部,将所述机器人具备的马达的扭矩信息和所述机器人的位置信息的至少一方作为状态变量观测;以及

学习部,基于所述状态变量学习所述动作参数。

9.根据权利要求8所述的控制装置,其特征在于,

所述位置信息基于所述机器人具备的惯性传感器的输出和配置于所述机器人的外部的位置检测部的输出的至少一方被算出。

10.根据权利要求8或9所述的控制装置,其特征在于,

所述学习部决定基于所述状态变量使所述动作参数变化的行为,使所述动作参数最佳化。

11.根据权利要求10所述的控制装置,其特征在于,

所述学习部基于所述机器人进行的作业的好坏与否,评价所述行为的回报。

12.根据权利要求11所述的控制装置,其特征在于,

所述学习部在以下的至少一种情况下将所述回报评价为正:所述作业的所需时间比基准短的情况,所述机器人的位置和目标位置的偏离量比基准小的情况,残留振动比基准小的情况,所述机器人的位置的超调比基准小的情况,以及发生声音比基准小的情况。

13.根据权利要求11或12所述的控制装置,其特征在于,

所述学习部在以下的至少一种情况下将所述回报评价为负:作业完成前丢下所述机器人把持的对象物的情况,作业完成前使作为所述机器人的作业对象的对象物的一部分分离的情况,以及作为所述机器人的作业对象的对象物破损的情况。

14.根据权利要求11至13中任一项所述的控制装置,其特征在于,

所述算出部通过反复进行所述状态变量的观测、对应于该状态变量的所述行为的决定和通过该行为得到的所述回报的评价,使所述动作参数最佳化。

15.根据权利要求10至14中任一项所述的控制装置,其特征在于,

在末端执行器被设置于所述机器人的状态下观测所述状态变量,

在所述末端执行器被设置于所述机器人的状态下执行所述行为。

16.根据权利要求15所述的控制装置,其特征在于,

在所述末端执行器把持对象物的状态下观测所述状态变量,

在所述末端执行器把持对象物的状态下执行所述行为。

17.根据权利要求1至16中任一项所述的控制装置,其特征在于,

所述算出部基于所述机器人的不同的多个所述动作算出多个所述动作共用的所述动作参数。

18.一种机器人,其特征在于,所述机器人通过权利要求1至17中任一项所述的控制装置来控制。

19.一种机器人系统,其特征在于,具备:

权利要求1至17中任一项所述的控制装置;以及

通过所述控制装置控制的所述机器人。

控制装置、机器人以及机器人系统

技术领域

[0001] 本发明涉及控制装置、机器人以及机器人系统。

背景技术

[0002] 为了使机器人进行作业而需要各种设定，以往，人为地进行各种设定。但是，为了进行该设定需要高度的技术，难度较高。因此，开发出了从远程支持控制参数的调整的技术(例如，专利文献1)。

[0003] 此外，以往，已知为了使工作机械的工具校正的频率最佳化而利用机械学习的技术(专利文献2)。

[0004] 现有技术文献

[0005] 专利文献1：日本专利特许第4739556号公报

[0006] 专利文献2：日本专利特许第5969676号公报。

发明内容

[0007] 即使利用以往的技术，考虑到机器人中的各种动作设定动作参数是困难的。因此，对充分挖掘出机器人的性能的动作参数进行设定依旧是困难的。

[0008] 为了解决上述课题的至少一个，控制装置具备：算出部，使用机械学习算出与机器人的动作相关的动作参数；以及控制部，基于算出的所述动作参数控制所述机器人。根据该结构，能够以比人为地决定的动作参数高的概率来算出进行高性能的动作的动作参数。

[0009] 进一步地，也可以构成为，动作参数包含用于控制机器人具备的马达的伺服增益。根据该结构，能够自动地调整较难通过人为的调整进行适当的设定的、用于控制马达的伺服增益。

[0010] 进一步地，也可以构成为，动作参数包含机器人的加减速特性。根据该结构，能够自动地调整较难通过人为的调整进行适当的设定的加减速特性。

[0011] 进一步地，也可以构成为，动作参数包含用于基于机器人具备的惯性传感器进行控制的伺服增益。根据该结构，能够自动地调整较难通过人为的调整进行适当的设定的、用于基于惯性传感器进行控制的伺服增益。

[0012] 动作参数不包含机器人的动作的起始点以及终点。根据该结构，能够防止从机器人预定的起始点以及终点偏离并进行利用者不打算进行的动作。

[0013] 进一步地，也可以构成为，动作参数不包含对机器人示教过的示教位置。根据该结构，能够防止从机器人示教过的位置偏离并进行利用者不打算进行的动作。

[0014] 进一步地，也可以构成为，动作参数包含机器人在通过阻抗控制进行动作时的阻抗参数。根据该结构，能够容易地算出挖掘出机器人的性能的阻抗参数。

[0015] 进一步地，也可以构成为，算出部包含：状态观测部，将机器人具备的马达的扭矩信息和机器人的位置信息的至少一方作为状态变量观测；以及学习部，基于状态变量学习动作参数。根据该结构，能够容易地算出进行高性能的动作的动作参数。

[0016] 进一步地，也可以构成为，位置信息基于机器人具备的惯性传感器的输出和配置于机器人的外部的位置检测部的输出的至少一方被算出。根据该惯性传感器，能够基于在机器人中通用使用的传感器算出位置信息。配置于机器人的外部的位置检测部不被机器人的动作影响而能够算出位置信息。

[0017] 进一步地，也可以构成为，学习部决定基于状态变量使动作参数变化的行为，使动作参数最佳化。根据该结构，能够以使成为对应于机器人的使用环境的动作参数的方式进行最佳化。

[0018] 进一步地，也可以构成为，学习部基于对象物的检测结果的好坏与否，评价行为的回报。根据该结构，能够执行提高对象物的检测精度的学习。

[0019] 进一步地，也可以构成为，学习部基于机器人进行的作业的好坏与否，评价行为的回报。根据该结构，能够以提高机器人的作业的品质的方式使参数最佳化。

[0020] 进一步地，也可以构成为，学习部在以下的至少一种情况下将所述回报评价为正：作业的所需时间比基准短的情况，机器人的位置和目标位置的偏离量比基准小的情况，残留振动比基准小的情况，机器人的位置的超调比基准小的情况，以及发生声音比基准小的情况。

[0021] 根据在作业的所需时间比基准短的情况下将回报评价为正的结构，能够容易地算出在短的时间使机器人作业的动作参数。根据在机器人的位置和目标位置的偏离量比基准小的情况下将回报评价为正的结构，能够容易地算出使机器人向目标位置正确地移动的动作参数。

[0022] 根据在残留振动比基准小的情况下将回报评价为正的结构，能够容易地算出使基于机器人的动作的振动产生的可能性较低的动作参数。根据在机器人的位置的超调比基准小的情况下将回报评价为正的结构，能够容易地算出机器人超调的可能性较低的动作参数。根据在发生声音比基准小的情况下将回报评价为正的结构，能够容易地算出使机器人产生异常的可能性较低的动作参数。

[0023] 进一步地，也可以构成为，学习部在以下的至少一种情况下将所述回报评价为负：作业完成前丢下机器人把持的对象物的情况，作业完成前使作为机器人的作业对象的对象物的一部分分离的情况，在作为机器人的作业对象的对象物破损的情况。根据在作业完成前丢下机器人把持有的对象物的情况下将回报评价为负的结构，能够容易地算出不使对象物丢下而使作业完成的可能性较高的动作参数。

[0024] 根据在作业完成前使机器人的作业对象即对象物的一部分分离的情况下将回报评价为负的结构，能够容易地算出不使对象物分离而使作业完成的可能性较高的动作参数。根据在机器人的作业对象即对象物破损了的情况下将回报评价为负的结构，能够容易地算出不使对象物破损的可能性较低的动作参数。

[0025] 进一步地，也可以构成为，算出部通过反复进行状态变量的观测、对应于该状态变量的行为的决定和通过该行为得到的回报的评价，使动作参数最佳化。根据该结构，能够使动作参数自动地最佳化。

[0026] 进一步地，也可以构成为，在末端执行器被设置于机器人的状态下观测状态变量，在末端执行器被设置于机器人的状态下执行行为。根据该结构，能够容易地算出适合于进行使用了末端执行器的动作的机器人的动作参数。

[0027] 进一步地，也可以构成为，在末端执行器把持对象物的状态下观测状态变量，在末端执行器把持对象物的状态下执行行为。根据该结构，能够容易地算出适合于通过末端执行器把持对象物并进行动作的机器人的动作参数。

[0028] 进一步地，也可以构成为，算出部基于机器人的不同的多个动作算出多个动作共用的动作参数。根据该结构，能够容易地算出能够应用于各种动作的通用的动作参数。

附图说明

[0029] 图1是机器人系统的立体图。

[0030] 图2是控制装置的功能框图。

[0031] 图3是示出参数的图。

[0032] 图4是示出加减速特性的图。

[0033] 图5是拾取处理的流程图。

[0034] 图6是与算出部关联的结构的框图。

[0035] 图7是示出学习光学参数时的例子的图。

[0036] 图8是示出多层神经网络的例子的图。

[0037] 图9是学习处理的流程图。

[0038] 图10是示出学习动作参数时的例子的图。

[0039] 图11是示出学习力控制参数时的例子的图。

[0040] 附图标记说明：

[0041] 1～3 机器人；20 光学系统；21 拍摄部；22 照明部；23 夹钳；40 控制装置；41 算出部；41a 状态观测部；41b 学习部；42 检测部；43 控制部；43a 位置控制部；43b 力控制部；43c 接触判定部；43d 伺服；44 存储部；44a 参数；44b 机器人程序；44c 模板数据；44d 行为信息；44e 学习信息。

具体实施方式

[0042] 以下，参照附图按以下的顺序对本发明的实施方式进行说明。另外，在各图中在对应的结构要素标注同一符号，并省略重复的说明。

[0043] (1)机器人系统的结构：

[0044] (2)机器人的控制：

[0045] (3)拾取处理：

[0046] (4)学习处理：

[0047] (4-1)光学参数的学习：

[0048] (4-2)光学参数的学习例：

[0049] (4-3)动作参数的学习：

[0050] (4-4)动作参数的学习例：

[0051] (4-5)力控制参数的学习：

[0052] (4-6)力控制参数的学习例：

[0053] (5)其他实施方式：

[0054] (1)机器人系统的结构：

[0055] 图1是示出由本发明的一实施方式的控制装置控制的机器人的立体图。如图1所示，作为本发明的一实施例的机器人系统具备机器人1～3。机器人1～3是具备末端执行器的六轴机器人，在机器人1～3安装有不同的末端执行器。即，在机器人1安装有拍摄部21，在机器人2安装有照明部22，在机器人3安装有夹钳23。另外，在此，将拍摄部21以及照明部22称为光学系统。

[0056] 机器人1～3由控制装置40控制。控制装置40通过线缆与机器人1～3能够通信地连接。另外，也可以将控制装置40的结构要素设于机器人1。此外，也可以由多个装置构成控制装置40(例如，将后述的学习部和控制部设于不同的装置等)。此外，控制装置40能够通过线缆或无线通信连接未图示的示教装置。示教装置可以是专用的计算机，也可以是安装有用于对机器人1进行示教的程序的通用的计算机。进一步地，也可以将控制装置40和示教装置一体地构成。

[0057] 机器人1～3是将各种末端执行器安装于臂而使用的单臂机器人，在本实施方式中，在机器人1～3中臂、轴的结构是同等的。在图1中在机器人3中标注有说明臂、轴的结构的符号。如在机器人3中所示的那样，机器人1～3具备基台T、六个臂部件A1～A6和六个关节J1～J6。基台T固定于作业台。基台T和六个臂部件A1～A6通过关节J1～J6连结。臂部件A1～A6和末端执行器是可动部，这些可动部进行动作，从而机器人1～3能够进行各种作业。

[0058] 在本实施方式中，关节J2、J3、J5是弯曲关节，关节J1、J4、J6是扭转关节。在臂A之中最末端侧的臂部件A6装配力觉传感器P和末端执行器。机器人1～3通过使六轴的臂驱动而能够在可动范围内将末端执行器配置在任意位置，并设为任意姿势(角度)。

[0059] 机器人3具备的末端执行器是夹钳23，能够把持对象物W。机器人2具备的末端执行器是照明部22，能够在照射范围照射光。机器人1具备的末端执行器是拍摄部21，能够拍摄视线内的图像。在本实施方式中，将相对于机器人1～3具备的末端执行器相对被固定的位置定义为工具中心点(TCP)。TCP的位置为末端执行器的基准的位置，TCP为原点，定义相对于末端执行器相对被固定的三维正交坐标系即TCP坐标系。

[0060] 力觉传感器P是六轴力检测器。力觉传感器P检测与传感器坐标系中相互正交的三个检测轴平行的力的大小和绕该三个检测轴的扭矩的大小，所述传感器坐标系是以力觉传感器上的点为原点的三维正交坐标系。另外，在本实施例中以六轴机器人为例，但机器人的形态可以是各种形态，机器人1～3的形态也可以不同。此外，也可以在关节J6以外的关节J1～J5的任一个以上设有作为力检测器的力觉传感器。

[0061] 在将规定设置有机器人1～3的空间的坐标系称为机器人坐标系时，机器人坐标系是由在水平面上相互正交的x轴、y轴和将铅垂朝向上设为正方向的z轴规定的三维正交坐标系(参照图1)。z轴中的负的方向大致与重力方向一致。此外，由Rx表示绕x轴的旋转角，由Ry表示绕y轴的旋转角，由Rz表示绕z轴的旋转角。能够通过x、y、z方向的位置来表现三维空间中的任意位置，能够通过Rx、Ry、Rz方向的旋转角来表现三维空间中的任意姿势。以下，在标记为位置的情况下，也可能意思是姿势。此外，在标记为力的情况下，也可能意思是扭矩。

[0062] 另外，在本实施方式中，能够执行对作用于机器人的力进行控制的力控制，在力控制中，以使作用于任意点的该作用力成为目标力的方式进行控制。在三维正交坐标系即力控制坐标系中定义作用于各种部位的力。目标力(包含扭矩)能够通过以由力控制坐标系所表现的力的作用点为起点的向量来表现，在进行后述的学习以前，目标力向量的起点是力

控制坐标系的原点，作用力的方向与力控制坐标系的一轴方向一致。但是，在进行了后述的学习的情况下，目标力向量的起点可以与力控制坐标系的原点不同，并且目标力向量的方向可以与力控制坐标系的轴方向不同。

[0063] 在本实施方式中预先定义各种坐标系的关系，各种坐标系中的坐标值能够相互转换。即，TCP坐标系、传感器坐标系、机器人坐标系、力控制坐标系中的位置、向量能够相互转换。在此，为了简单而进行控制装置40在机器人坐标系中控制TCP的位置以及作用于TCP的作用力的说明，但由于能够在各种坐标系中定义机器人1～3的位置、作用于机器人1～3的力并能够相互转换，因此也可以在任何坐标系中定义位置、力并进行控制。当然，除了在此所阐述的坐标系以外，也可以定义其他坐标系(例如固定于对象物的对象物坐标系等)并能够转换。

[0064] (2)机器人的控制：

[0065] 机器人1是通过进行示教而能够进行各种作业的通用机器人，如图2所示，具备作为致动器的马达M1～M6和作为传感器的编码器E1～E6。对臂进行控制意思是控制马达M1～M6。与各个关节J1～J6对应地设有马达M1～M6和编码器E1～E6，编码器E1～E6检测马达M1～M6的旋转角度。此外，在各马达M1～M6连接有供给电力的电源线，在各电源线设有电流计。因此，控制装置40能够测量供给到各马达M1～M6的电流。

[0066] 控制装置40具备计算机等的硬件资源和存储于存储部44的各种软件资源，能够执行程序。在本实施方式中控制装置40作为算出部41、检测部42、控制部43发挥功能。另外，硬件资源能够采用各种结构，可以构成为由CPU、RAM、ROM等结构，也可以由ASIC等构成。

[0067] 在本实施方式中检测部42能够执行检测对象物的处理，控制部43能够驱动机器人1～3的臂。检测部42与构成光学系统20的拍摄部21和照明部22连接。检测部42能够控制拍摄部21，获取由拍摄部21具备的拍摄传感器拍摄到的图像。此外，检测部42能够控制照明部22，使输出光的亮度变化。

[0068] 当从拍摄部21输出图像时，检测部42基于拍摄图像进行模板匹配处理，并进行检测对象物的位置(位置姿势)的处理。即，检测部42基于存储于存储部44的模板数据44c执行模板匹配处理。模板数据44c是多个位置姿势的每个的模板。因此，若通过ID等将位置姿势相对模板数据44c建立对应，则能够根据匹配的模板数据44c的种类来确定从检测部42观察到的对象物的位置姿势。

[0069] 具体地，检测部42将多个位置姿势的每个的模板数据44c依次作为处理对象，一边使模板数据44c的大小变化一边与拍摄到的图像进行比较。然后，检测部42将模板数据44c与图像的差分在阈值以下的图像作为对象物的图像检出。

[0070] 当检测对象物的图像时，检测部42基于与预先决定的坐标系的关系匹配的模板数据44c的大小来确定对象物的位置姿势。即，从模板数据44c的大小判明拍摄部21与对象物的光轴方向的距离，从在图像内检测到的对象物的位置判明与光轴垂直的方向的位置。

[0071] 因此，例如，若将拍摄部21的拍摄传感器的光轴和拍摄平面上的两轴定义为与TCP坐标系的各轴平行的情况下，检测部42基于模板数据44c的大小和模板数据44c与图像匹配的位置而能够在TCP坐标系中确定对象物的位置。此外，检测部42基于匹配的模板数据44c的ID而能够确定TCP坐标系中的对象物的姿势。因此，检测部42能够利用上述的坐标系的对应关系来确定例如机器人坐标系的任意坐标系中的对象物的位置姿势。

[0072] 另外，模板匹配处理只要是用于确定对象物的位置姿势的处理即可，能够采用各种处理。例如，模板数据44c与图像的差分可以通过灰度值的差分来评价，也可以通过图像的特征(例如，图像的梯度等)的差分来评价。

[0073] 检测部42参照参数进行该模板匹配处理。即，在存储部44存储有各种参数44a，在该参数44a中包含有与检测部42的检测相关的参数。图3是示出参数44a的例子的图。在图3所示的例子中，参数44a包含有光学参数、动作参数和力控制参数。

[0074] 光学参数是与检测部42的检测相关的参数。动作参数和力控制参数是控制机器人1～3时的参数，后述详情。光学参数包含与拍摄部21相关的拍摄部参数、与照明部22相关的照明部参数和与针对由拍摄部21拍摄到的对象物的图像的图像处理相关的图像处理参数。

[0075] 在图3中，示出了这些参数的例子。即，在拍摄对象物时将配置拍摄部21的位置作为拍摄部的位置来定义，并包含于拍摄部参数。此外，拍摄部21具备能够调整曝光时间和光圈的机构，拍摄对象物时的曝光时间以及光圈的值包含于拍摄部参数。另外，拍摄部的位置可以通过各种方法来记述，例如能够采用通过机器人坐标系来记述拍摄部21的TCP的位置的结构等。

[0076] 检测部42参照拍摄部参数将拍摄部21的位置向后述的位置控制部43a传送。其结果，位置控制部43a生成目标位置L_t，基于该目标位置L_t控制机器人1。此外，检测部42参照拍摄部参数设定拍摄部21的曝光时间和光圈。其结果，在拍摄部21中成为根据该曝光时间和光圈进行拍摄的状态。

[0077] 此外，在拍摄对象物时将配置照明部22的位置作为照明部的位置来定义，并包含于照明部参数。此外，照明部22具备能够调整亮度的机构，拍摄对象物时的亮度的值包含于照明部参数。另外，照明部的位置也可以通过各种方法来记述，例如能够采用通过机器人坐标系来记述照明部22的TCP的位置的结构等。

[0078] 检测部42参照照明部参数将照明部22的位置向后述的位置控制部43a传送。其结果，位置控制部43a生成目标位置L_t，基于该目标位置L_t控制机器人2。此外，检测部42参照照明部参数设定照明部22的亮度。其结果，在照明部22中成为输出该亮度的光的状态。

[0079] 检测部42在对由拍摄部21拍摄到的图像应用模板匹配处理时参照图像处理参数。即，在图像处理参数中包含有示出在执行模板匹配处理时的处理顺序的图像处理序列。此外，在本实施方式中，模板匹配处理中的阈值是可变的，当前的模板匹配处理的阈值包含于图像处理参数。进一步地，检测部42在对模板数据44c和图像进行比较之前能够执行各种处理。在图3中，作为各种处理例示出平滑化处理和锐化处理，各自的强度包含于图像处理参数。

[0080] 当从拍摄部21输出图像时，检测部42基于图像处理序列决定图像处理的顺序(包含是否执行)，按该顺序执行平滑化处理、锐化处理等的图像处理。此时，检测部42以记述于图像处理参数的强度来执行平滑化处理、锐化处理等的图像处理。此外，在执行包含于图像处理序列的比较(模板数据44c与图像的比较)时，基于图像处理参数示出的阈值来进行比较。

[0081] 另外，如以上那样检测部42能够基于光学参数确定拍摄部21、照明部22的位置，并使机器人1、机器人2动作，但驱动机器人1以及机器人2时的位置也可以通过后述的动作参数、力控制参数赋予。

[0082] 在本实施方式中，控制部43具备位置控制部43a、力控制部43b、接触判定部43c、伺服43d。此外，在控制部43中，使马达M1～M6的旋转角度的组合与机器人坐标系中的TCP的位置的对应关系U1存储于未图示的存储介质，定义坐标系的对应关系U2，并存储于未图示的存储介质。因此，控制部43、后述的算出部41能够基于对应关系U2将任意坐标系中的向量转换为其他坐标系中的向量。例如，控制部43、算出部41能够基于力觉传感器P的输出获取传感器坐标系中的朝机器人1～3的作用力，并转换为作用于机器人坐标系中的TCP位置的力。此外，控制部43、算出部41能够将通过力控制坐标系所表现的目标力转换为机器人坐标系中的TCP的位置中的目标力。当然，也可以使对应关系U1、U2存储于存储部44。

[0083] 控制部43通过驱动臂而能够控制与机器人1～3一起移动的各种部位的位置、作用于各种部位的力，位置的控制主要通过位置控制部43a来执行，力的控制主要通过力控制部43b来执行。伺服43d能够执行伺服控制，并执行使编码器E1～E6的输出示出的马达M1～M6的旋转角度D_a与控制目标即目标角度D_t一致的反馈控制。即，伺服43d能够执行PID控制，所述PID控制使伺服增益Kpp、Kpi、Kpd作用于旋转角度D_a与目标角度D_t的偏差、该偏差的积分、该偏差的微分。

[0084] 进一步地，伺服43d能够执行PID控制，所述PID控制使伺服增益Kvp、Kvi、Kvd作用于该伺服增益Kpp、Kpi、Kpd所作用的输出与旋转角度D_a的微分的偏差、该偏差的积分、该偏差的微分。能够对各个马达M1～M6执行该伺服43d的控制。因此，关于机器人1～3具备的六轴的各个能够执行各伺服增益。另外，在本实施方式中，控制部43能够向伺服43d输出控制信号，使伺服增益Kpp、Kpi、Kpd、Kvp、Kvi、Kvd变化。

[0085] 在存储部44中除了上述的参数44a还存储用于控制机器人1～3的机器人程序44b。在本实施方式中，参数44a以及机器人程序44b通过示教而生成，并存储于存储部44，可以通过后述的算出部41修正。另外，机器人程序44b主要示出机器人1～3实施的作业的序列(工序的顺序)，通过预先定义的指令的组合来记述。此外，参数44a主要是为了实现各工序而设为需要的具体的值，作为各指令的自变量来记述。

[0086] 在用于控制机器人1～3的参数44a中除了上述的光学参数之外还包含动作参数和力控制参数。动作参数是与机器人1～3的动作相关的参数，在本实施方式中是在位置控制时参照的参数。即，在本实施方式中，使一系列的作业划分为多个工序，通过示教而生成实施各工序时的参数44a。在动作参数中包含有示出该多个工序中的起始点和终点的参数。该起始点和终点可以在各种坐标系中定义，在本实施方式中，在机器人坐标系中定义控制对象的机器人的TCP的起始点以及终点。即，定义关于机器人坐标系的各轴的平移位置和旋转位置。

[0087] 此外，在动作参数中包含有多个工序中的TCP的加减速特性。加减速特性示出机器人1～3的TCP在从各工序的起始点至终点移动时的期间和该期间内的各时刻中的TCP的速度。图4是示出该加减速特性的例子的图，在从起始点中的TCP的移动开始时刻t_1至TCP到达终点的时刻t_4的期间内的各时刻定义了TCP的速度V。此外，在本实施方式中在加减速特性中包含定速期间。

[0088] 定速期间是时刻t_2～t_3的期间，在该期间内速度是一定的。此外，在该期间的前后TCP加速或减速。即，在时刻t_1～t_2期间TCP加速，在时刻t_3～t_4期间TCP减速。该加减速特性也可以在各种坐标系中定义，在本实施方式中是关于控制对象的机器人的TCP的速度，在机

器人坐标系中定义。即，定义关于机器人坐标系的各轴的平移速度和旋转速度(角速度)。

[0089] 进一步地，在动作参数中包含有伺服增益Kpp、Kpi、Kpd、Kvp、Kvi、Kvd。即，控制部43能够以使成为作为动作参数而记述的值的方式向伺服43d输出控制信号，调整伺服增益Kpp、Kpi、Kpd、Kvp、Kvi、Kvd。在本实施方式中该伺服增益是上述的每个工序的值，但也可以通过后述的学习等设为更短的每个期间的值。

[0090] 力控制参数是与机器人1～3的力控制相关的参数，在本实施方式中是在力控制时参照的参数。起始点、终点、加减速特性、伺服增益是与动作参数同样的参数，关于机器人坐标系的三轴的平移和旋转定义起始点、终点、加减速特性。此外，关于各个马达M1～M6定义伺服增益。但是，在力控制的情况下，有时也不对起始点以及终点之中的至少一部分进行定义(设为任意)。例如，在以使作用于某一方向的力成为零的方式进行冲撞回避、模仿控制的情况下，有时也不对该方向上的起始点以及终点进行定义，而以使该方向的力为零的方式定义使位置可以任意变化的状态。

[0091] 此外，在力控制参数中包含有示出力控制坐标系的信息。力控制坐标系是用于定义力控制的目标力的坐标系，在进行后述的学习之前目标力向量的起点是原点，一轴朝向目标力向量的方向。即，在示教中在定义力控制中的各种目标力时，示教各作业的各工序中的目标力的作用点。例如，在使对象物的一点与其他物体抵接，在两者的接触点在从对象物使一定的目标力作用于其他物体的状态下使对象物的朝向变化的情况下，对象物与其他物体的接触的点成为目标力的作用点，定义以该作用点为原点的力控制坐标系。因此，在力控制参数中，使如下信息包含在参数中：用于确定以力控制的目标力进行作用的点为原点并使一轴朝向目标力的方向的坐标系的信息，即用于确定力控制坐标系的信息。另外，该参数能够进行各种定义，例如能够通过示出力控制坐标系和其他坐标系(机器人坐标系等)的关系的数据来定义。

[0092] 进一步地，在力控制参数中包含有目标力。目标力是在各种作业中作为应该作用于任意点的力而被示教的力，在力控制坐标系中进行定义。即，使示出目标力的目标力向量作为目标力向量的起点和来自起点的六轴成分(三轴平移力、三轴扭矩)来定义，并由力控制坐标系来表现。另外，若利用力控制坐标系和其他坐标系的关系，则能够将该目标力转换为任意坐标系、例如机器人坐标系中的向量。

[0093] 进一步地，在力控制参数中包含有阻抗参数。即，在本实施方式中力控制部43b实施的力控制是阻抗控制。阻抗控制是通过马达M1～M6来实现虚拟的机械性阻抗的控制。这时，使TCP虚拟地具有的质量作为虚拟惯性系数m来定义，使TCP虚拟地受到的粘性阻力作为虚拟粘性系数d来定义，使TCP虚拟地受到的弹性力的弹簧常数作为虚拟弹性系数k来定义。阻抗参数是这些m、d、k，并且关于相对于机器人坐标系的各轴的平移和旋转进行定义。在本实施方式中该力控制坐标系、目标力、阻抗参数是上述的每个工序的值，但也可以通过后述的学习等设为更短的每个期间的值。

[0094] 在本实施方式中，使一系列的作业划分为多个工序，通过示教而生成实施各工序的机器人程序44b，位置控制部43a将机器人程序44b示出的各工序进一步地细化分为每个微小时间ΔT的微小工序。然后，位置控制部43a基于参数44a生成每个微小工序的目标位置L_t。力控制部43b基于参数44a获取一系列的作业的各工序中的目标力f_{Lt}。

[0095] 即，位置控制部43a参照动作参数或力控制参数示出的起始点、终点、加减速特性，

将在以该加减速特性从起始点移动至终点的情况下(在是姿势的情况下为姿势变化的情况下)的每个微小工序的TCP的位置作为目标位置L_t生成。力控制部43b参照关于各工序的力控制参数示出的目标力,基于力控制坐标系与机器人坐标系的对应关系U2将该目标力转换为机器人坐标系中的目标力f_{Lt}。该目标力f_{Lt}可以作为作用于任意点的力来转换,但在此,由于后述的作用力被作为作用于TCP的力来表现,因此为了通过运动方程式解析该作用力和目标力f_{Lt}而使目标力f_{Lt}作为转换为TCP的位置中的力来进行说明。当然,有时也根据工序不定义目标力f_{Lt},在该情况下,进行不伴随力控制的位置控制。

[0096] 另外,在此文字L表示规定机器人坐标系的轴的方向(x、y、z、Rx、Ry、Rz)之中的任意一个方向。此外,L也表示L方向的位置。例如,在L=x的情况下,使在机器人坐标系中所设定的目标位置的x方向成分标记为$L_t=xt$,使目标力的x方向成分标记为$f_{Lt}=fxt$。

[0097] 为了执行位置控制、力控制,控制部43能够获取机器人1～3的状态。即,控制部43能够获取马达M1～M6的旋转角度D_a,基于对应关系U1将该旋转角度D_a转换为机器人坐标系中的TCP的位置L(x、y、z、Rx、Ry、Rz)。此外,控制部43能够参照对应关系U2,基于TCP的位置L和力觉传感器P的检测值以及位置,将现实中作用于力觉传感器P的力转换为作用于TCP的作用力f_L并在机器人坐标系中进行确定。

[0098] 即,在传感器坐标系中定义作用于力觉传感器P的力。因此,控制部43基于机器人坐标系中的TCP的位置L、对应关系U2和力觉传感器P的检测值,在机器人坐标系中确定作用于TCP的作用力f_L。此外,作用于机器人的扭矩能够由从作用力f_L和工具接触点(末端执行器与工件的接触点)至力觉传感器P的距离来算出,并被作为未图示的f_L扭矩成分来确定。另外,控制部43对作用力f_L进行重力补偿。重力补偿是从作用力f_L除去重力成分的处理。重力补偿例如能够通过预先调查针对TCP的每个姿势作用于TCP的作用力f_L的重力成分,并从作用力f_L减去该重力成分等来实现。

[0099] 当确定作用于TCP的重力以外的作用力f_L和应该作用于TCP的目标力f_{Lt}时,力控制部43b在对象物等的物体存在于TCP并可以使力作用于该TCP的状态下,获取基于阻抗控制的校正量ΔL(以后,称为力来源校正量ΔL)。即,力控制部43b参照参数44a获取目标力f_{Lt}和阻抗参数m、d、k,代入运动方程式(1)而获取力来源校正量ΔL。另外,该力来源校正量ΔL意思是,在TCP受到机械性阻抗的情况下,为了解除目标力f_{Lt}与力偏差$\Delta f_L(t)$,TCP应该移动的位置L的大小。

[0100] [数学式1]

[0101] $$m\Delta\ddot{L}(t)+d\Delta\dot{L}(t)+k\Delta L(t)=\Delta f_L(t) \quad \cdots(1)$$

[0102] (1)式的左边由如下项构成:将虚拟惯性系数m与TCP的位置L的二阶微分值相乘后的第一项,将虚拟粘性系数d与TCP的位置L的微分值相乘后的第二项,和将虚拟弹性系数k与TCP的位置L相乘后的第三项。(1)式的右边由从目标力f_{Lt}减去现实的作用力f_L后的力偏差$\Delta f_L(t)$构成。(1)式中的微分意思是时间的微分。

[0103] 当得到力来源校正量ΔL时,控制部43基于对应关系U1将规定机器人坐标系的各轴的方向的动作位置转换为各马达M1～M6的目标的旋转角度、即目标角度D_t。伺服43d通过从目标角度D_t减去马达M1～M6的现实的旋转角度、即编码器E1～E6的输出(旋转角度D_a)而算出驱动位置偏差$D_e=(D_t-D_a)$。伺服43d参照参数44a获取伺服增益Kpp、Kpi、Kpd、Kvp、Kvi、Kvd,通过将伺服增益Kpp、Kpi、Kpd与驱动位置偏差D_e相乘后的值加上将伺服增益Kvp、Kvi、

Kvd与和作为现实的旋转角度D_a的时间微分值的驱动速度之差即驱动速度偏差相乘后的值，从而导出控制量Dc。关于各个马达M1～M6确定控制量Dc，通过各马达M1～M6的控制量Dc控制各个马达M1～M6。控制部43控制马达M1～M6的信号是被PWM(Pulse Width Modulation:脉冲宽度调制)调制后的信号。

[0104] 如以上那样，将基于运动方程式从目标力f_{Lt}导出控制量Dc并控制马达M1～M6的模式称为力控制模式。此外，控制部43在末端执行器等的结构要素未从对象物W受到力的非接触状态的工序中不进行力控制，而通过由线形运算从目标位置导出的旋转角度来控制马达M1～M6。将通过由线形运算从目标位置导出的旋转角度来控制马达M1～M6的模式称为位置控制模式。进一步地，控制部43将由线形运算从目标位置导出的旋转角度和将目标力代入运动方程式而导出的旋转角度例如通过线型结合来进行统一，通过统一后的旋转角度来控制马达M1～M6的混合模式也能够控制机器人1。通过机器人程序44b预先决定这些模式。

[0105] 在通过位置控制模式或混合模式进行控制的情况下，位置控制部43a获取每个微小工序的目标位置L_t。当得到每个微小工序的目标位置L_t时，控制部43基于对应关系U1将规定机器人坐标系的各轴的方向的动作位置转换为各马达M1～M6的目标的旋转角度即目标角度D_t。伺服43d参照参数44a获取伺服增益Kpp、Kpi、Kpd、Kvp、Kvi、Kvd，基于目标角度D_t导出控制量Dc。关于各个马达M1～M6确定控制量Dc，通过各马达M1～M6的控制量Dc控制各个马达M1～M6。其结果，在各工序中，TCP经由每个微小工序的目标位置L_t，按照加减速特性从起始点移动至终点。

[0106] 另外，在混合模式中，控制部43将力来源校正量ΔL与每个微小工序的目标位置L_t相加，从而确定动作位置(L_t+ΔL)，基于该动作位置获取目标角度D_t，并获取控制量Dc。

[0107] 接触判定部43c执行判定机器人1～3在作业中是否与未设想到的物体已接触的功能。在本实施方式中，接触判定部43c获取各个机器人1～3具备的力觉传感器P的输出，在输出超过预先决定的基准值的情况下判定为机器人1～3在作业中与未设想到的物体已接触。在该情况下，可以进行各种处理，在本实施方式中接触判定部43c将机器人1～3的控制量Dc设为零而使机器人1～3停止。另外，停止时的控制量可以是各种控制量，也可以是以取消之前的控制量Dc的控制量而使机器人1～3动作的结构等。

[0108] (3)拾取处理：

[0109] 接下来，说明以上的结构中的机器人1～3的动作。在此，以通过机器人2的照明部22进行照明、通过机器人3的夹钳23拾取由机器人1的拍摄部21拍摄到的对象物W的作业为例进行说明。当然，机器人1～3的作业并不限定于拾取作业，此外也能够应用于各种作业(例如，螺丝紧固作业、插入作业、通过钻头进行的开孔作业、去毛刺作业、研磨作业、组装作业、产品检查作业等)。拾取处理通过上述的指令所记述的机器人控制程序而由检测部42以及控制部43执行的处理来实现。在本实施方式中拾取处理在对象物W配置于作业台的状态下执行。

[0110] 图5是示出拾取处理的流程图的例子的图。当开始拾取处理时，检测部42获取拍摄部21拍摄到的图像(步骤S100)。即，检测部42参照参数44a确定照明部22的位置，对位置控制部43a传送该位置。其结果，位置控制部43a执行以当前的照明部22的位置为起始点、以参数44a示出的照明部22的位置为终点的位置控制，使照明部22向参数44a示出的照明部的位置移动。接下来，检测部42参照参数44a确定照明部22的亮度，控制照明部22将照明的亮度

设定为该亮度。

[0111] 进一步地，检测部42参照参数44a确定拍摄部21的位置，对位置控制部43a传送该位置。其结果，位置控制部43a执行以当前的拍摄部21的位置为起始点、以参数44a示出的拍摄部21的位置为终点的位置控制，使拍摄部21向参数44a示出的拍摄部的位置移动。接下来，检测部42参照参数44a确定拍摄部21的曝光时间以及光圈，控制拍摄部21将曝光时间以及光圈设定为该曝光时间以及光圈。当曝光时间以及光圈的设定完成时，拍摄部21拍摄图像，并对检测部42输出。检测部42获取该图像。

[0112] 接下来，检测部42基于图像来判定对象物的检测是否已成功(步骤S105)。即，检测部42参照参数44a来确定图像处理序列，以参数44a示出的强度来执行该图像处理序列示出的各处理。此外，检测部42参照模板数据44c，将模板数据44c和图像的差分与阈值进行比较，在差分在阈值以下的情况下，判定为对象物的检测已成功。

[0113] 在步骤S105中，在未判定为对象物的检测已成功的情况下，检测部42使模板数据44c与图像的相对位置、或模板数据44c的大小的至少一方变化，反复进行步骤S100之后的处理。另一方面，在步骤S105中，在判定为对象物的检测已成功的情况下，控制部43确定控制目标(步骤S110)。

[0114] 本例中的拾取处理是与检测部42检测到的对象物W的位置姿势配合使机器人3的夹钳23移动、使姿势变化、通过夹钳23拾取对象物W、将对象物W搬运至规定的位置并从夹钳23放下对象物W的作业。因此，位置控制部43a以及力控制部43b基于机器人程序44b确定构成一系列的作业的多个工序。

[0115] 成为控制目标的特定对象的工序是在各工序之中未处理且按时序先存在的工序。在成为控制目标的特定对象的工序是力控制模式的工序的情况下，力控制部43b参照参数44a的力控制参数，获取力控制坐标系、目标力。力控制部43b基于力控制坐标系将该目标力转换为机器人坐标系的目标力f_{Lt}。此外，力控制部43b将力觉传感器P的输出转换为作用于TCP的作用力f_L。进一步地，力控制部43b参照参数44a的力控制参数，基于阻抗参数m、d、k，获取力来源校正量ΔL作为控制目标。

[0116] 在成为控制目标的特定对象的工序是位置控制模式的情况下，位置控制部43a将该工序细化分为微小工序。然后，位置控制部43a参照参数44a的动作参数，基于起始点、终点以及加减速特性，获取每个微小工序的目标位置L_t作为控制目标。在成为控制目标的特定对象的工序是混合模式的情况下，位置控制部43a将该工序细化分为微小工序，参照参数44a的力控制参数，基于起始点、终点以及加减速特性，获取每个微小工序的目标位置L_t，基于力控制坐标系、目标力f_{Lt}、阻抗参数、作用力f_L来获取力来源校正量ΔL。这些目标位置L_t以及力来源校正量ΔL是控制目标。

[0117] 当确定控制目标时，伺服43d以当前的控制目标来控制机器人3(步骤S115)。即，在当前的工序是力控制模式或混合模式的工序的情况下，伺服43d参照参数44a的力控制参数，基于伺服增益，确定与控制目标对应的控制量Dc，控制各个马达M1～M6。在当前的工序是位置控制模式的工序的情况下，伺服43d参照参数44a的动作参数，基于伺服增益，确定与控制目标对应的控制量Dc，控制各个马达M1～M6。

[0118] 接下来，控制部43判定当前的工序是否已结束(步骤S120)。该判定可以通过各种结束判定条件来执行，若为位置控制，则例如可列举TCP到达目标位置、在目标位置中TCP进

行了整定等。若为力控制，则例如可列举从作用力与目标力一致的状态变化为作用力在指定的大小以上或指定的大小以下、TCP变为指定的范围外等。前者例如可列举拾取作业中的对象物的把持动作的完成、把持解除动作的完成等。后者例如可列举在通过钻头进行的对象物的贯通作业中钻头已贯通的情况等。

[0119] 当然，此外在推定为各工序已失败的情况下，也可以判定为工序已结束。但是，在该情况下，优选进行作业的中止、中断。作为用于判定工序的失败的结束判定条件，例如可列举TCP的移动速度、加速度超过上限值的情况、发生了超时的情况等。是否满足了结束判定条件，也可以利用各种传感器、力觉传感器P、拍摄部21、其他传感器等。

[0120] 在步骤S120中，在未判定为当前的工序已结束的情况下，控制部43在微小时间ΔT后关于下一微小工序执行步骤S115之后的处理。即，在当前的工序是位置控制模式或混合模式的情况下，位置控制部43a将下一微小工序中的目标位置L_t作为控制目标来控制机器人3。此外，在当前的工序是力控制模式或混合模式的情况下，力控制部43b再次基于力觉传感器P的输出来获取作用力f_L，将基于最新的作用力f_L确定的力来源校正量ΔL作为控制目标来控制机器人3。

[0121] 在步骤S120中，在判定为当前的工序已结束的情况下，控制部43判定作业是否已结束(步骤S125)。即，在步骤S120中在判定为已结束的工序是最终工序的情况下，控制部43判定为作业已结束。在步骤S125中在未判定为作业已结束的情况下，控制部43将作业序列的下一工序变更为当前的工序(步骤S130)，执行步骤S110之后的处理。在步骤S125中在判定为已结束的情况下，控制部43判定为作业已结束，结束拾取处理。

[0122] (4)学习处理：

[0123] 本实施方式的控制装置40如以上那样能够基于参数44a控制机器人1～3。在上述的实施方式中，参数44a通过示教而生成，但通过人为的示教使参数44a最佳化是困难的。

[0124] 例如，在基于检测部42的对象物W的检测中，即使是相同对象物W，若光学参数不同则对象物W的位置、图像内的对象物的图像、位置、在对象物W产生的影子等各种各样的要素也可能变化。因此，当使光学参数变化时，基于检测部42的对象物W的检测精度可能变化。然后，在使光学参数变化了的情况下对象物W的检测精度会怎样变化未必是明确的。

[0125] 此外，在机器人1～3的控制利用动作参数、力控制参数，如机器人1～3那样具有多个自由度(可动轴)的机器人能够以极其多的模式进行动作。然后，在机器人1～3中，需要以使振动、异响、超调等不优选的动作不发生的方式决定模式。进一步地，在作为末端执行器而安装各种装置的情况下，由于机器人1～3的重心可能变化，因此最佳的动作参数、力控制参数也可能变化。然后，在使动作参数、力控制参数变化了的情况下机器人1～3的动作会怎样变化未必是明确的。

[0126] 进一步地，在机器人1～3中在进行力控制的情况下利用力控制参数，在机器人1～3中在实施的各作业中，在使力控制参数变化了的情况下，机器人1～3的动作会怎样变化未必是明确的。例如，在全部的作业工序中很难推定在什么样的方向上怎样的阻抗参数是最佳的。因此，为了提高检测部42的检测精度或者挖掘出机器人1～3的潜在的性能而需要进行极其多的试错。

[0127] 但是，由于人为地进行极其多的试错是困难的，因此人为地实现如下状态是困难的：对象物W的检测精度充分高而推定该检测精度大致到达上限的状态、挖掘出机器人1～3

的潜在的性能的状态(使所需时间、耗电等的性能进一步的提升较困难的状态)。此外,为了进行参数44a的调整而需要熟知基于参数44a的变化的检测精度的变化、机器人1～3的动作的变化的操作员,不熟知的操作员进行参数44a的调整是困难的。此外,总是需要熟知的操作员的系统是非常不方便的。

[0128] 因此,在本实施方式中,具备不进行人为的参数44a的决定作业而用于自动地决定参数44a的结构。另外,根据本实施方式,能够实现通过参数44a的稍微的变更而推定检测精度不会更提升(推定为检测精度为极其大)的状态、通过参数44a的稍微的变更而推定机器人1～3的性能不会高性能化(推定为性能为极其大)的状态。在本实施方式中,将这些状态称为最佳化后的状态。

[0129] 在本实施方式中,控制部40为了参数44a的自动的决定而具备算出部41。在本实施方式中,算出部41能够使用机械学习算出光学参数、动作参数和力控制参数。图6是示出算出部41的结构的图,是省略图2所示的结构的一部分并示出了算出部41的详情的图。另外,图6所示的存储部44是与图2所示的存储部44相同的存储介质,在各图中省略了所存储的信息的一部分的图示。

[0130] 算出部41具备观测状态变量的状态观测部41a和基于观测到的状态变量学习参数44a的学习部41b。在本实施方式中,状态观测部41a将通过使参数44a变化而产生的结果作为状态变量来观测。因此,状态观测部41a能够获取伺服43d的控制结果、编码器E1～E6的值、力觉传感器P的输出和检测部42获取的图像作为状态变量。

[0131] 具体地,状态观测部41a将向马达M1～M6供给的电流值作为伺服43d的控制结果来观测。该电流值相当于由马达M1～M6输出的扭矩。使编码器E1～E6的输出基于对应关系U1转换为机器人坐标中的TCP的位置。因此,若是机器人1则状态观测部41a能够观测拍摄部21的位置,若是机器人2则能够观测照明部22的位置,若是机器人3则能够观测夹钳23的位置。

[0132] 使力觉传感器P的输出基于对应关系U2转换为作用于机器人坐标系中的TCP的作用力。因此,状态观测部41a能够将朝机器人1～3的作用力作为状态变量来观测。检测部42获取的图像是由拍摄部21拍摄到的图像,状态观测部41a能够将该图像作为状态变量来观测。状态观测部41a能够根据学习对象的参数44a适当选择观测对象的状态变量。

[0133] 学习部41b只要能够通过学习使参数44a最佳化即可,在本实施方式中,通过强化学习使参数44a最佳化。具体地,学习部41b决定基于状态变量使参数44a变化的行为,并执行该行为。若根据该行为后的状态评价回报,则判明该行为的行为价值。因此,算出部41通过反复进行状态变量的观测、对应于该状态变量的行为的决定和由该行为得到的回报的评价而使参数44a最佳化。

[0134] 在本实施方式中,算出部41能够从参数44a之中选择学习对象的参数并进行学习。在本实施方式中,能够独立地执行光学参数的学习、动作参数的学习和力控制参数的学习的各个。

[0135] (4-1)光学参数的学习:

[0136] 图7是依据由代理(agent)和环境构成的强化学习的模式来说明光学参数的学习例的图。图7所示的代理相当于根据预先决定的策略选择行为a的功能,通过学习部41b实现。环境相当于基于代理所选择的行为a和当前的状态s来决定下一状态s'并基于行为a、状态s和状态s'来决定即时回报r的功能,通过状态观测部41a以及学习部41b实现。

[0137] 在本实施方式中，采用如下Q学习：学习部41b通过预先决定的策略选择行为a，并反复进行状态观测部41a进行状态的变更的操作，从而算出某一状态s下的某一行为a的行为价值函数Q(s,a)。即，在本例中，通过下述的式(2)更新行为价值函数。然后，在使行为价值函数Q(s,a)合理地收敛的情况下，视为使该行为价值函数Q(s,a)最大化的行为a是最佳的行为，视为示出该行为a的参数44a是被最佳化后的参数。

[0138] [数学式2]

[0139] $$Q(s_t,a_t)\leftarrow Q(s_t,a_t)+\alpha\left(r_{t+1}+\gamma \max_{a'} Q(s_{t+1},a')-Q(s_t,a_t)\right)\cdots(2)$$

[0140] 在此，行为价值函数Q(s,a)是在状态s下在采取了行为a的情况下在将来得到的收益(在本例中是折扣回报总和)的期望值。回报是r，状态s、行为a、回报r的下标t是示出按时序反复进行的试行过程中的一次份的步骤的编号(称为试行编号)，当在行为决定后状态变化时使试行编号递增。因此，式(2)内的回报r_{t+1}是在状态s_t下选择行为a_t而状态变为s_{t+1}的情况下得到的回报。α是学习率，γ是折扣率。此外，a'是在状态s_{t+1}下可以采取的行为a_{t+1}之中使行为价值函数$Q(s_{t+1},a_{t+1})$最大化的行为，$\max_{a'}Q(s_{t+1},a')$是通过选择了行为a'而被最大化后的行为价值函数。

[0141] 在光学参数的学习中，使光学参数变化相当于行为的决定，使示出学习对象的参数和可以采取的行为的行为信息44d预先记录于存储部44。即，作为学习对象在该行为信息44d记述的光学参数成为学习对象。在图7中，示出了光学参数之中的拍摄部参数、照明部参数和图像处理参数的一部分成为学习对象的例子。

[0142] 具体地，在拍摄部参数之中拍摄部21的x坐标、y坐标成为学习对象。因此，在该例子中z坐标、相对于xyz轴的旋转(姿势)不会成为学习对象，而拍摄部21为朝向放置对象物W的作业台的状态，同时在拍摄部21的x-y平面内的移动是学习对象。当然，其他拍摄部参数例如拍摄部21的姿势、z坐标、曝光时间、光圈也可以是学习对象。

[0143] 此外，在图7所示的例子中，在照明部参数之中，照明部22的x坐标、y坐标以及照明部的亮度成为学习对象。因此，在该例子中，z坐标、相对于xyz轴的旋转(姿势)不会成为学习对象，而照明部22为朝向放置对象物W的作业台的状态，同时在照明部22的x-y平面内的移动是学习对象。当然，其他照明部参数例如照明部22的姿势、z坐标也可以是学习对象。

[0144] 进一步地，在图7所示的例子中，在图像处理参数之中，平滑化处理的强度、锐化处理的强度和模板匹配处理的阈值成为学习对象。因此，在该例子中，图像处理序列不会成为学习对象，而相对于由拍摄部21拍摄到的图像的图像处理的顺序不会变化(当然，也能够采用图像处理序列是学习对象的实施方式)。

[0145] 在图7所示的例子中，在行为中存在使值增加一定值的行为和使值减少一定值的行为。因此，在图7所示的全部八个参数中可以采取的行为是全部十六个(行为a1～行为a16)。由于行为信息44d示出了学习对象的参数和可以采取的行为，因此若是图7所示的例子，则使图8所示的八个参数作为学习对象记述于行为信息44d。此外，将用于确定各行为的信息(行为的ID、各行为中的增减量等)记述于行为信息44d。

[0146] 在图7所示的例子中，基于对象物W的检测的成功与否来确定回报。即，学习部41b在使光学参数作为行为a变化了之后通过该光学参数使机器人1、2动作，并通过检测部42获取拍摄部21拍摄到的图像。然后，学习部41b基于该光学参数执行模板匹配处理，判定对象物W的检测是否已成功。进一步地，学习部41b根据检测的成功与否决定行为a、状态s、s'的

回报。该回报只要基于对象物W的检测的成功与否来决定即可，例如，能够采用对检测的成功赋予正(例如+1)的、对检测的失败赋予负(例如-1)的回报的结构等。根据该结构，能够以提高对象物的检测精度的方式进行最佳化。

[0147] 在进行了作为行为a的参数的变化之后使机器人1、2动作，状态观测部41a观测状态，从而能够确定在当前的状态s下在采用了行为a的情况下的下一状态s'。另外，在本例的光学参数的学习中机器人3不动作。在图7所示的例子中，在状态变量中包含有拍摄部21的x坐标、y坐标、照明部22的x坐标、y坐标、照明部22的亮度、平滑化处理的强度、锐化处理的强度、模板匹配的阈值和由拍摄部21拍摄到的图像。

[0148] 因此，在该例子中，状态观测部41a在执行过行为a之后基于U1来转换机器人1的编码器E1～E6的输出并观测拍摄部21的x坐标以及y坐标。此外，状态观测部41a在执行过行为a之后基于U1来转换机器人2的编码器E1～E6的输出并观测照明部22的x坐标以及y坐标。

[0149] 在本实施方式中，视为照明部22的亮度能够通过参数44a无误差地进行调整(或视为误差没有影响)，状态观测部41a获取包含于参数44a的照明部的亮度而视为观测到状态变量。当然，照明部22的亮度也可以通过传感器等来实际测量，也可以基于拍摄部21拍摄到的图像(例如，通过平均灰度值等)来观测。关于平滑化处理的强度、锐化处理的强度、模板匹配的阈值，状态观测部41a也参照参数44a获取当前的值，视为观测到状态变量。

[0150] 进一步地，在状态观测部41a中，获取拍摄部21拍摄到的、检测部42获取到的图像作为状态变量(图7所示的粗框)。即，状态观测部41a将拍摄部21拍摄到的图像(也可以是可能存在对象物的感兴趣区域等的图像)的每个像素的灰度值作为状态变量来观测。拍摄部的x坐标等是行为同时是作为观测对象的状态，但拍摄部21拍摄到的图像不是行为。因此，在该意义上，拍摄到的图像是可以进行从光学参数的变化直接进行推定较困难的变化的状态变量。此外，由于检测部42基于该图像检测对象物，因此该图像是可以对检测的成功与否直接带来影响的状态变量。因此，作为状态变量，能够通过观测该图像而进行人为地改善较困难的参数的改善，并以有效地提高检测部42的检测精度的方式使光学参数最佳化。

[0151] (4-2)光学参数的学习例：

[0152] 接下来，说明光学参数的学习例。将示出在学习的过程中参照的变量、函数的信息作为学习信息44e存储于存储部44。即，算出部41采用了如下结构：通过反复进行状态变量的观测、对应于该状态变量的行为的决定和由该行为得到的回报的评价而使行为价值函数Q(s,a)收敛。因此，在本例中，使在学习的过程中状态变量、行为和回报的时序的值依次不断记录于学习信息44e。

[0153] 行为价值函数Q(s,a)可以通过各种方法算出，也可以基于多次试行来算出，但在本实施方式中，采用了近似性地算出行为价值函数Q(s,a)的方法，即DQN(Deep Q-Network)。在DQN中，使用多层神经网络推定行为价值函数Q(s,a)。在本例中，采用了将状态s作为输入、将可以选择的行为的数N个行为价值函数Q(s,a)的值作为输出的多层神经网络。

[0154] 图8是示意性地示出在本例中所采用的多层神经网络的图。在图8中，多层神经网络将M个(M是2以上的整数)状态变量作为输入、将N个(N是2以上的整数)的行为价值函数Q的值作为输出。例如，若是图7所示的例子，则拍摄部的x坐标～模板匹配的阈值的八个状态变量与拍摄到的图像的像素数之和是M个，使M个状态变量的值向多层神经网络输入。在图8

中，将试行编号t中的M个状态作为s_{1t}～s_{Mt}示出。

[0155] N个是可以选择的行为a的数量，多层神经网络的输出是在所输入的状态s下在选择了特定的行为a的情况下的行为价值函数Q的值。在图8中，将在试行编号t中可以选择的各个行为a_{1t}～a_{Nt}中的行为价值函数Q作为$Q(s_t,a_{1t})$～$Q(s_t,a_{Nt})$示出。包含于该Q的s_t是代表所输入的状态s_{1t}～s_{Mt}而示出的文字。若是图7所示的例子，则由于能够选择十六个行为，因此N＝16。当然，行为a的内容、数量(N的值)、状态s的内容、数量(M的值)也可以根据试行编号t而变化。

[0156] 图8所示的多层神经网络是在各层的各节点中执行相对于之前的层的输入(在第一层中是状态s)的权值w的相乘和偏置b的相加，并根据需要执行得到经过了激活函数的输出(成为下一层的输入)的运算的模式。在本例中，存在P个(P是1以上的整数)层DL，在各层中存在多个节点。

[0157] 图8所示的多层神经网络通过各层中的权值、偏置b、激活函数和层的顺序等来确定。因此，在本实施方式中，将用于确定该多层神经网络的参数(为了从输入得到输出而需要的信息)作为学习信息44e记录于存储部44。另外，在学习时，不断更新在用于确定多层神经网络的参数之中可变的值(例如，权值w和偏置b)。在此，将在学习的过程中可能变化的多层神经网络的参数标记为θ。当使用该θ时，也能够将上述的行为价值函数$Q(s_t,a_{1t})$～$Q(s_t,a_{Nt})$标记为$Q(s_t,a_{1t};\theta_t)$～$Q(s_t,a_{Nt};\theta_t)$。

[0158] 接下来，沿着图9所示的流程图来说明学习处理的顺序。光学参数的学习处理可以在机器人1、2的运用过程中实施，也可以在实际运用之前事先执行学习处理。在此，按照在实际运用之前事先执行学习处理的结构(当使示出多层神经网络的θ最佳化时，是保存该信息并在下一次之后的运用中利用的结构)说明学习处理。

[0159] 当开始学习处理时，算出部41将学习信息44e初始化(步骤S200)。即，算出部41确定在开始学习时参照的θ的初始值。初始值可以通过各种方法来决定，在过去未进行学习的情况下，可以使任意值、随机值等为θ的初始值，也可以准备模拟机器人1、2、拍摄部21、照明部22的光学特性的模拟环境并基于该环境将所学习或所推定的θ作为初始值。

[0160] 在过去进行了学习的情况下，将该学习完毕的θ作为初始值采用。此外，在过去进行了关于类似的对象的学习的情况下，也可以使该学习中的θ设为初始值。过去的学习可以是用户使用机器人1、2来进行，也可以是机器人1、2的制造商在机器人1、2的销售前进行。在该情况下，也可以构成为制造商根据对象物、作业的种类预备好多个初始值的组，在用户学习时选择初始值。当决定θ的初始值时，使该初始值作为当前的θ的值存储于学习信息44e。

[0161] 接下来，算出部41将参数初始化(步骤S205)。在此，由于光学参数是学习对象，因此算出部41将光学参数初始化。即，算出部41通过对应关系U1转换机器人1的编码器E1～E6的输出，将拍摄部21的位置作为初始值来设定。此外，算出部41将预先决定的初始的曝光时间(在过去进行了学习的情况下是最新的曝光时间)作为拍摄部21的曝光时间的初始值来设定。进一步地，算出部41向拍摄部21输出控制信号，将当前的光圈的值作为初始值来设定。

[0162] 进一步地，算出部41通过对应关系U1转换机器人2的编码器E1～E6的输出，将照明部22的位置作为初始值来设定。此外，算出部41将预先决定的初始的亮度(在过去进行了学习的情况下是最新的亮度)作为照明部22的亮度的初始值来设定。进一步地，算出部41关于

平滑化处理的强度、锐化处理的强度、模板匹配的阈值、图像处理序列而设定预先决定的初始值(在过去进行了学习的情况下是最新的值)。使被初始化后的参数作为当前的参数44a存储于存储部44。

[0163] 接下来,状态观测部41a观测状态变量(步骤S210)。即,控制部43参照参数44a以及机器人程序44b控制机器人1、2。检测部42基于在控制后的状态下拍摄部21拍摄到的图像而执行对象物W的检测处理(相当于上述的步骤S100、S105)。这之后,状态观测部41a基于U1来转换机器人1的编码器E1～E6的输出,并观测拍摄部21的x坐标以及y坐标。此外,状态观测部41a基于U1来转换机器人2的编码器E1～E6的输出,并观测照明部22的x坐标以及y坐标。进一步地,状态观测部41a参照参数44a获取应该设定于照明部22的亮度,视为观测到状态变量。

[0164] 进一步地,状态观测部41a关于平滑化处理的强度、锐化处理的强度、模板匹配的阈值也参照参数44a获取当前的值,视为观测到状态变量。进一步地,在状态观测部41a中,获取拍摄部21拍摄到的、检测部42获取到的图像,并获取各像素的灰度值作为状态变量。

[0165] 接下来,学习部41b算出行为价值(步骤S215)。即,学习部41b参照学习信息44e获取θ,向示出学习信息44e的多层神经网络输入最新的状态变量,算出N个行为价值函数$Q(s_t,a_{1t};\theta_t)$～$Q(s_t,a_{Nt};\theta_t)$。

[0166] 另外,最新的状态变量在初次的执行时是步骤S210的观测结果,在第二次之后的执行时是步骤S225的观测结果。此外,试行编号t在初次的执行时为零,在第二次之后的执行时为1以上的值。在过去未实施学习处理的情况下,由于未使学习信息44e示出的θ最佳化,因此作为行为值函数Q的值可能变为不正确的值,但通过步骤S215以后的处理的反复,从而使行为值函数Q逐渐不断最佳化。此外,在步骤S215以后的处理的反复中,能够使状态s、行为a、回报r与各试行编号t建立对应地存储于存储部44,在任意定时(timing)参照。

[0167] 接下来,学习部41b选择行为并执行(步骤S220)。在本实施方式中,进行视为使行为价值函数Q(s,a)最大化的行为a是最佳的行为的处理。因此,学习部41b确定在步骤S215中算出来的N个行为价值函数$Q(s_t,a_{1t};\theta_t)$～$Q(s_t,a_{Nt};\theta_t)$的值之中最大的值。然后,学习部41b选择赋予了最大的值的行为。例如,若在N个行为价值函数$Q(s_t,a_{1t};\theta_t)$～$Q(s_t,a_{Nt};\theta_t)$之中$Q(s_t,a_{1t};\theta_t)$是最大值,则学习部41b选择行为a_{Nt}。

[0168] 当选择行为时,学习部41b使与该行为对应的参数44a变化。例如,在图7所示的例子中,在选择了使拍摄部的x坐标增加一定值的行为a1的情况下,学习部41b在光学参数的拍摄部参数示出的拍摄部的位置中使x坐标增加一定值。当进行参数44a的变化时,控制部43参照该参数44a控制机器人1、2。检测部42基于在控制后的状态下拍摄部21拍摄到的图像而执行对象物W的检测处理。

[0169] 接下来,状态观测部41a观测状态变量(步骤S225)。即,状态观测部41a进行与步骤S210中的状态变量的观测同样的处理,作为状态变量而获取拍摄部21的x坐标以及y坐标、照明部22的x坐标以及y坐标、应该设定于照明部22的亮度、平滑化处理的强度、锐化处理的强度、模板匹配的阈值、拍摄部21拍摄到的图像的各像素的灰度值。另外,在当前的试行编号是t的情况下(所选择的行为是a_t的情况下),在步骤S225中获取的状态s是s_{t+1}。

[0170] 接下来,学习部41b评价回报(步骤S230)。在本例中,基于对象物W的检测的成功与否来决定回报。因此,学习部41b从检测部42获取对象物的检测结果的成功与否(步骤S105

的成功与否)，若检测成功则获取既定量的正的回报，若检测失败则获取既定量的负的回报。另外，在当前的试行编号是t的情况下，在步骤S230中获取的回报r是r_{t+1}。

[0171] 在本实施方式中，谋求式(2)所示的行为价值函数Q的更新，但为了使行为价值函数Q不断适当地更新而必须将示出行为价值函数Q的多层神经网络最佳化(使θ最佳化)。为了通过图8所示的多层神经网络使行为价值函数Q适当地输出而需要成为该输出的目标的教师数据。即，通过以使多层神经网络的输出与目标的误差最小化的方式改善θ，从而可期待使多层神经网络最佳化。

[0172] 但是，在本实施方式中，在学习未完成的阶段没有行为价值函数Q的认知，确定目标是困难的。因此，在本实施方式中，通过式(2)的第二项、使所谓的TD误差(Temporal Difference:时序差分)最小化的目标函数来实施示出多层神经网络的θ的改善。即，将$(r_{t+1}+\gamma \max_{a'}Q(s_{t+1},a';\theta_t))$作为目标，以使目标与$Q(s_t,a_t;\theta_t)$的误差最小化的方式学习θ。但是，由于目标$(r_{t+1}+\gamma \max_{a'}Q(s_{t+1},a';\theta_t))$包含有学习对象的θ，因此在本实施方式中经过某一程度的试行次数来固定目标(例如，通过最后学习到的θ(初次学习时是θ的初始值)来固定)。在本实施方式中，预先决定固定目标的试行次数即既定次数。

[0173] 由于在这样的前提下进行学习，因此当在步骤S230中评价回报时，学习部41算出目标函数(步骤S235)。即，学习部41b算出用于评价试行的各自中的TD误差的目标函数(例如，与TD误差的2次方的期望值成比例的函数、TD误差的2次方的总和等)。另外，由于TD误差在固定有目标的状态下被算出，因此当将被固定后的目标标记为$(r_{t+1}+\gamma \max_{a'}Q(s_{t+1},a';\theta_-))$时，TD误差是$(r_{t+1}+\gamma \max_{a'}Q(s_{t+1},a';\theta_-)-Q(s_t,a_t;\theta_t))$。在该TD误差的式中回报$r_{t+1}$是根据行为$a_t$在步骤S230中得到的回报。

[0174] 此外，$\max_{a'}Q(s_{t+1},a';\theta_-)$是将根据行为$a_t$在步骤S225中算出的状态$s_{t+1}$作为通过被固定后的$\theta_-$确定的多层神经网络的输入的情况下得到的输出之中的最大值。$Q(s_t,a_t;\theta_t)$是在将选择了行为a_t的前一状态s_t作为通过试行编号t的阶段的θ_t确定的多层神经网络的输入的情况下得到的输出之中、对应于行为a_t的输出的值。

[0175] 当算出目标函数时，学习部41b判定学习是否已结束(步骤S240)。在本实施方式中，预先决定用于判定TD误差是否充分小的阈值，在目标函数在阈值以下的情况下，学习部41b判定为学习已结束。

[0176] 在步骤S240中在未判定为学习已结束的情况下，学习部41b更新行为价值(步骤S245)。即，学习部41b基于TD误差的θ的偏微分来确定用于使目标函数减小的θ的变化，并使θ变化。当然，在此，能够通过各种方法使θ变化，例如，能够采用RMSProp等的梯度下降法。此外，也可以适当实施学习率等的调整。根据以上的处理，能够以使行为价值函数Q接近目标的方式使θ变化。

[0177] 但是，在本实施方式中，如上述那样，由于目标被固定，因此学习部41b进一步地进行是否更新目标的判定。具体地，学习部41b判定是否进行了既定次数的试行(步骤S250)，在步骤S250中，在判定为进行了既定次数的试行的情况下，学习部41b更新目标(步骤S255)。即，学习部41b将在算出目标时参照的θ更新为最新的θ。这之后，学习部41b反复进行步骤S215之后的处理。另一方面，在步骤S250中，若未判定为进行了既定次数的试行的情况下，则学习部41b跳过步骤S255反复进行步骤S215之后的处理。

[0178] 在步骤S240中在判定为学习已结束的情况下，学习部41b更新学习信息44e(步骤

S260)。即，学习部41b将通过学习得到的θ作为在机器人1、2的作业、检测部42的检测时应该参照的θ在学习信息44e记录。在记录有包含该θ的学习信息44e的情况下，如步骤S100～S105那样在进行机器人1、2的作业时，检测部42基于参数44a进行对象物的检测处理。然后，至检测部42的检测成功为止，在反复进行拍摄部21的拍摄的工序中，反复进行基于状态观测部41a的当前的状态的观测和基于学习部41b的行为的选择。当然，这时，学习部41b选择在将状态作为输入而被算出来的输出Q(s,a)之中赋予最大值的行为a。然后，在选择了行为a的情况下，以使成为相当于进行了行为a的状态的值的方式更新参数44a。

[0179] 根据以上的结构，检测部42能够一边选择使行为价值函数Q最大化的行为a一边执行对象物的检测处理。反复进行了多数试行的结果，该行为价值函数Q通过上述的处理而被最佳化。然后，该试行通过算出部41自动进行，能够容易地执行不能够人为地实施的多数试行。因此，根据本实施方式，能够以比人为地决定的光学参数高的概率高精度地检测对象物。

[0180] 进一步地，在本实施方式中，由于检测部42构成为检测对象物的位置姿势，因此根据本实施方式能够高精度地检测对象物的位置姿势。进一步地，根据本实施方式，能够基于最佳化后的行为价值函数Q算出光学参数即拍摄部参数。因此，能够以提高对象物的检测精度的方式调整拍摄部21。进一步地，根据本实施方式，能够基于最佳化后的行为价值函数Q算出光学参数即照明部参数。因此，能够以提高对象物的检测精度的方式调整照明部22。

[0181] 进一步地，根据本实施方式，能够基于最佳化后的行为价值函数Q算出光学参数即图像处理参数。因此，能够执行提高对象物的检测精度的图像处理。进一步地，根据本实施方式，由于自动使行为价值函数Q最佳化，因此能够容易地算出高精度地检测对象物的光学参数。此外，由于自动地进行行为价值函数Q的最佳化，因此也能够自动地进行最佳的光学参数的算出。

[0182] 进一步地，在本实施方式中，学习部41b决定基于作为状态变量的图像使光学参数变化的行为，使光学参数最佳化。因此，能够在通过照明部22进行了照明的实际环境下基于由拍摄部21实际上拍摄到的图像使光学参数最佳化。因此，能够以成为对应于机器人1、2的使用环境的光学参数的方式进行最佳化。

[0183] 在本实施方式中，拍摄部21的位置以及照明部22的位置包含于行为中，基于该行为使行为价值函数Q最佳化，从而能够使与拍摄部21的位置以及照明部22的位置相关的参数44a最佳化。因此，在学习后，至少使拍摄部21与照明部22的相对位置关系理想化。此外，若使对象物W放置于作业台的固定位置或大致被固定后的位置，则也能够考虑在学习后使拍摄部21与照明部22的在机器人坐标系中的位置理想化。进一步地，在本实施方式中，使由拍摄部21拍摄到的图像作为状态被观测。因此，根据本实施方式，使与各种图像的状态对应的拍摄部21的位置、照明部22的位置理想化。

[0184] (4-3)动作参数的学习：

[0185] 在动作参数的学习中，也能够选择学习对象的参数，在此，说明其一例。图10是通过与图7同样的模式对动作参数的学习例进行了说明的图。本例也基于式(2)使行为价值函数Q(s,a)最佳化。因此，视为使最佳化后的行为价值函数Q(s,a)最大化的行为a是最佳的行为，视为示出该行为a的参数44a是被最佳化后的参数。

[0186] 在动作参数的学习中，使动作参数变化也相当于行为的决定，使示出学习对象的

参数和可以采取的行为的行为信息44d预先记录于存储部44。即，作为学习对象记述于该行为信息44d的动作参数成为学习对象。在图10中，机器人3中的动作参数之中的伺服增益和加减速特性是学习对象，动作的起始点以及终点未成为学习对象。另外，动作的起始点以及终点是示教位置，但在本实施方式中未示教其他位置。因此，在本实施方式中，构成为不包含对机器人3示教过的位置。

[0187] 具体地，关于各个马达M1～M6，定义动作参数之中的伺服增益Kpp、Kpi、Kpd、Kvp、Kvi、Kvd，关于六轴的各个能够增减。因此，在本实施方式中，能够针对每一轴使六个伺服增益的各个增加或减少，关于增加可以选择三十六个行为，关于减少也可以选择三十六个行为，共计可以选择七十二个行为(行为a1～a72)。

[0188] 另一方面，动作参数之中的加减速特性是如图4所示的那样的特性，关于各个马达M1～M6(关于六轴)而被定义。在本实施方式中加减速特性能够使加速域中的加速度、减速域中的加速度、车速比零大的期间的长度(图4所示的t_4)变化。另外，在本实施方式中加速域、减速域中的曲线由加速度的增减来定义，例如，增减后的加速度示出曲线中央的倾斜，按照预先决定的规则使该中央的周围的曲线变化。当然，此外加减速特性的调整法也能够采用各种方法。

[0189] 无论怎样，在本实施方式中，针对每一轴能够通过三个要素(加速域、减速域、期间)调整加减速特性，能够使对应于各要素的数值(加速度、期间长度)增加或减少。因此，关于增加可以选择十八个行为，关于减少也可以选择十八个行为，共计可以选择三十六个行为(行为a73～a108)。在本实施方式中使与通过以上那样预先定义的行为的选择项对应的参数作为学习对象记述于行为信息44d。此外，将用于确定各行为的信息(行为的ID、各行为中的增减量等)记述于行为信息44d。

[0190] 在图10所示的例子中，基于机器人3所进行的作业的好坏与否评价回报。即，学习部41b在使动作参数作为行为a变化了之后，通过该动作参数使机器人3动作，并执行拾取通过检测部42检测到的对象物的作业。进一步地，学习部41b观测作业的好坏与否，评价作业的好坏与否。然后，学习部41b根据作业的好坏与否决定行为a、状态s、s'的回报。

[0191] 另外，作业的好坏与否不仅包含作业的成功与否(拾取的成功与否等)还包含作业的品质。具体地，学习部41b基于未图示的计时电路获取从作业的开始至结束(从步骤S110的开始至在步骤S125中判定为结束)的所需时间。然后，学习部41b在作业所需时间比基准短的情况下赋予正(例如+1)的回报，在作业所需时间比基准长的情况下赋予负(例如-1)的回报。另外，基准可以根据各种要素来确定，例如，可以是上次的作业所需时间，可以是过去的最短所需时间，并且也可以是预先决定的时间。

[0192] 进一步地，学习部41b在作业的各工序中基于U1来转换机器人3的编码器E1～E6的输出并获取夹钳23的位置。然后，学习部41b获取各工序的目标位置(终点)与工序结束时的夹钳23的位置的偏离量，在夹钳23的位置与目标位置的偏离量比基准小的情况下赋予正的回报，在比基准大的情况下赋予负的回报。另外，基准可以根据各种要素来确定，例如，可以是上次的偏离量，可以是过去的最短的偏离量，并且也可以是预先决定的偏离量。

[0193] 进一步地，学习部41b在经过进行整定以前的预定期间获取在作业的各工序中获取到的夹钳23的位置，并获取该期间中的振动强度。然后，学习部41b在该振动强度的程度比基准小的情况下赋予正的回报，在比基准大的情况下赋予负的回报。另外，基准可以根据

各种要素来确定，例如，可以是上次的振动强度的程度，可以是过去的最小的振动强度的程度，并且也可以是预先决定的振动强度的程度。振动强度的程度可以通过各种方法来确定，能够采用各种方法：自目标位置的背离的积分值、阈值以上的振动产生的期间等。另外，预定期间能够设为各种期间，若是从工序的起始点到终点的期间，则评价基于动作中的振动强度的回报，若使工序的尾期设为预定期间，则评价基于残留振动的强度的回报。

[0194] 进一步地，学习部41b在经过进行整定以前的预定期间获取在作业的各工序的尾期获取到的夹钳23的位置，并获取自该期间中的目标位置的背离的最大值作为超调量。然后，学习部41b在该超调量比基准小的情况下赋予正的回报，在比基准大的情况下赋予负的回报。另外，基准可以根据各种要素来确定，例如，可以是上次的超调量的程度，可以是过去的最小的超调量，并且也可以是预先决定的超调量。

[0195] 进一步地，在本实施方式中，在控制装置40、机器人1～3、作业台等的至少一处安装有声音采集装置，学习部41b获取示出在作业中声音采集装置获取到的声音的信息。然后，学习部41b在作业中的发生声音的大小比基准小的情况下赋予正的、在比基准大的情况下赋予负的回报。另外，基准可以根据各种要素来确定，例如，可以是上次的作业或工序的发生声音的大小的程度，可以是过去的发生声音的大小的最小值，并且也可以是预先决定的大小。此外，发生声音的大小可以通过声压的最大值来评价，也可以通过预定期间内的声压的统计值(平均值等)来评价，能够采用各种结构。

[0196] 在进行了作为行为a的参数的变化之后使机器人3动作，状态观测部41a观测状态，从而能够确定在当前的状态s下在采用了行为a的情况下的下一状态s'。另外，在基于机器人1、2的对象物的检测完成后，关于机器人3执行本例的动作参数的学习。

[0197] 在图10所示的例子中，在状态变量中包含有马达M1～M6的电流、编码器E1～E6的值、力觉传感器P的输出。因此，状态观测部41a能够观测向马达M1～M6供给的电流值作为伺服43d的控制结果。该电流值相当于由马达M1～M6输出的扭矩。此外，使编码器E1～E6的输出基于对应关系U1转换为机器人坐标系中的TCP的位置。因此，状态观测部41a观测机器人3具备的夹钳23的位置信息。

[0198] 力觉传感器P的输出通过积分而能够转换为机器人的位置。即，状态观测部41a基于对应关系U2在机器人坐标系中将朝TCP的作用力进行积分，从而获取TCP的位置。因此，在本实施方式中状态观测部41a也利用力觉传感器P的输出观测机器人3具备的夹钳23的位置信息。另外，状态可以通过各种方法来观测，也可以使不进行上述的转换的值(电流值、编码器、力觉传感器的输出值)作为状态来观测。

[0199] 状态观测部41a不是直接观测行为即伺服增益、加减速特性的调整结果，而是将调整的结果、将由机器人3得到的变化作为马达M1～M6的电流、编码器E1～E6的值、力觉传感器P的输出来观测。因此，变为间接观测行为的影响，在该意义上，本实施方式的状态变量是可以进行从动作参数的变化直接进行推定较困难的变化的状态变量。

[0200] 此外，马达M1～M6的电流、编码器E1～E6的值、力觉传感器P的输出直接示出了机器人3的动作，该动作直接示出了作业的好坏与否。因此，作为状态变量，观测马达M1～M6的电流、编码器E1～E6的值、力觉传感器P的输出，从而能够进行人为地改善较困难的参数的改善，并以有效地提高作业的品质的方式使动作参数最佳化。其结果，能够以比人为地决定的动作参数高的概率来算出进行高性能的动作的动作参数。

[0201] (4-4)动作参数的学习例：

[0202] 接下来，说明动作参数的学习例。使示出在学习的过程中参照的变量、函数的信息作为学习信息44e存储于存储部44。即，算出部41采用了如下结构：通过反复进行状态变量的观测、对应于该状态变量的行为的决定和由该行为得到的回报的评价而使行为价值函数Q(s,a)收敛。因此，在本例中，使在学习的过程中状态变量、行为和回报的时序的值依次不断记录于学习信息44e。

[0203] 另外，在本实施方式中，通过位置控制模式来执行动作参数的学习。为了执行位置控制模式中的学习，可以使仅通过位置控制模式构成的作业作为机器人3的机器人程序44b而生成，在使包含任意模式的作业作为机器人3的机器人程序44b而生成的状况下，也可以仅使用其中的位置控制模式进行学习。

[0204] 行为价值函数Q(s,a)可以通过各种方法来算出，也可以基于多次试行来算出，在此，对通过DQN使行为价值函数Q最佳化的例子进行说明。使利用于行为价值函数Q的最佳化的多层神经网络在上述的图8中示意性地示出。若是观测图10所示的那样的状态的本例，则由于机器人3中的马达M1～M6的电流、编码器E1～E6的值、力觉传感器P的输出(六轴的输出)为状态，因此状态s的数量M＝18。此外，若是可以选择图10所示的行为的本例，则由于能够选择一百零八个行为，因此N＝108。当然，行为a的内容或数量(N的值)、状态s的内容或数量(M的值)也可以根据试行编号t而变化。

[0205] 在本实施方式中，也将用于确定该多层神经网络的参数(为了从输入得到输出而需要的信息)作为学习信息44e记录于存储部44。在此，也将在学习的过程中可能变化的多层神经网络的参数标记为θ。当使用该θ时，也能够将上述的行为价值函数$Q(s_t,a_{1t})$～$Q(s_t,a_{Nt})$标记为$Q(s_t,a_{1t};\theta_t)$～$Q(s_t,a_{Nt};\theta_t)$。

[0206] 接下来，沿着图9所示的流程图来说明学习处理的顺序。动作参数的学习处理可以在机器人3的运用过程中实施，也可以在实际运用之前事先执行学习处理。在此，按照在实际运用之前事先执行学习处理的结构(当使示出多层神经网络的θ最佳化时，是保存该信息并在下一次之后的运用中利用的结构)说明学习处理。

[0207] 当开始学习处理时，算出部41将学习信息44e初始化(步骤S200)。即，算出部41确定在开始学习时参照的θ的初始值。初始值可以通过各种方法来决定，在过去未进行学习的情况下，可以使任意值、随机值等为θ的初始值，也可以准备模拟机器人3、对象物的模拟环境并基于该环境将所学习或所推定的θ作为初始值。

[0208] 在过去进行了学习的情况下，将该学习完毕的θ作为初始值采用。此外，在过去进行了关于类似的对象的学习的情况下，也可以使该学习中的θ设为初始值。过去的学习可以是用户使用机器人3来进行，也可以是机器人3的制造商在机器人3的销售前进行。在该情况下，也可以构成为制造商根据对象物、作业的种类预备好多个初始值的组，在用户学习时选择初始值。当决定θ的初始值时，使该初始值作为当前的θ的值存储于学习信息44e。

[0209] 接下来，算出部41将参数初始化(步骤S205)。在此，由于动作参数是学习对象，因此算出部41将动作参数初始化。即，若是未进行学习的状态，则算出部41将包含于通过示教而生成的参数44a的动作参数作为初始值来设定。若是过去进行了某种学习的状态，则算出部41将包含于在学习时最后利用的参数44a的动作参数作为初始值来设定。

[0210] 接下来，状态观测部41a观测状态变量(步骤S210)。即，控制部43参照参数44a以及

机器人程序44b控制机器人3(相当于上述的步骤S110～S130)。这之后,状态观测部41a观测向马达M1～M6供给的电流值。此外,状态观测部41a获取编码器E1～E6的输出,基于对应关系U1转换为机器人坐标系中的TCP的位置。进一步地,状态观测部41a将力觉传感器P的输出积分,获取TCP的位置。

[0211] 接下来,学习部41b算出行为价值(步骤S215)。即,学习部41b参照学习信息44e获取θ,向示出学习信息44e的多层神经网络输入最新的状态变量,算出N个行为价值函数$Q(s_t,a_{1t};\theta_t)$～$Q(s_t,a_{Nt};\theta_t)$。

[0212] 另外,最新的状态变量在初次的执行时是步骤S210的观测结果,在第二次之后的执行时是步骤S225的观测结果。此外,试行编号t在初次的执行时为零,在第二次之后的执行时为1以上的值。在过去未实施学习处理的情况下,由于未使学习信息44e示出的θ最佳化,因此作为行为值函数Q的值可能变为不正确的值,但通过步骤S215以后的处理的反复,从而使行为值函数Q逐渐不断最佳化。此外,在步骤S215以后的处理的反复中,能够使状态s、行为a、回报r与各试行编号t建立对应地存储于存储部44,在任意定时参照。

[0213] 接下来,学习部41b选择行为并执行(步骤S220)。在本实施方式中,进行视为使行为价值函数Q(s,a)最大化的行为a是最佳的行为的处理。因此,学习部41b确定在步骤S215中算出来的N个行为价值函数$Q(s_t,a_{1t};\theta_t)$～$Q(s_t,a_{Nt};\theta_t)$的值之中最大的值。然后,学习部41b选择赋予了最大的值的行为。例如,若在N个行为价值函数$Q(s_t,a_{1t};\theta_t)$～$Q(s_t,a_{Nt};\theta_t)$之中$Q(s_t,a_{1t};\theta_t)$是最大值,则学习部41b选择行为a_{Nt}。

[0214] 当选择行为时,学习部41b使与该行为对应的参数44a变化。例如,在图10所示的例子中,在选择了使马达M1的伺服增益Kpp增加一定值的行为a1的情况下,学习部41b在动作参数示出的马达M1的伺服增益Kpp的值增加一定值。当进行参数44a的变化时,控制部43参照该参数44a控制机器人3,使执行一系列的作业。另外,在本实施方式中,每当进行行为选择则执行一系列的作业,但也可以构成为每当进行行为选择则执行一系列的作业的一部分(构成为执行构成一系列的作业的多个工序的至少一工序)。

[0215] 接下来,状态观测部41a观测状态变量(步骤S225)。即,状态观测部41a进行与步骤S210中的状态变量的观测同样的处理,作为状态变量而获取向马达M1～M6供给的电流值、基于编码器E1～E6的输出确定的TCP的位置、基于力觉传感器P的输出确定的TCP的位置。另外,在当前的试行编号是t的情况下(所选择的行为是a_t的情况下),在步骤S225中获取的状态s是s_{t+1}。

[0216] 接下来,学习部41b评价回报(步骤S230)。即,学习部41b基于未图示的计时电路获取从作业的开始至结束的所需时间,在作业的所需时间比基准短的情况下获取正的回报,在作业的所需时间比基准长的情况下获取负的回报。进一步地,学习部41b获取作业的各工序的结束阶段中的夹钳23的位置,获取与各工序的目标位置的偏离量。然后,学习部41b在夹钳23的位置与目标位置的偏离量比基准小的情况下获取正的回报,在比基准大的情况下获取负的回报。在使一系列的作业由多个工序构成的情况下,可以获取各工序的回报之和,也可以获取统计值(平均值等)。

[0217] 进一步地,学习部41b基于在作业的各工序中获取到的夹钳23的位置来获取振动强度。然后,学习部41b在该振动强度的程度比基准小的情况下获取正的回报,在比基准大的情况下获取负的回报。在使一系列的作业由多个工序构成的情况下,可以获取各工序的

回报之和，也可以获取统计值(平均值等)。

[0218] 进一步地，学习部41b基于在作业的各工序的尾期获取到的夹钳23的位置来获取超调量。然后，学习部41b在使在该超调量比基准小的情况下获取正的回报、在比基准大的情况下获取负的回报的一系列的作业由多个工序构成的情况下，可以获取各工序的回报之和，也可以获取统计值(平均值等)。

[0219] 进一步地，学习部41b获取示出在作业中声音采集装置获取到的声音的信息。然后，学习部41b在作业中的发生声音的大小比基准小的情况下获取正的回报，在比基准大的情况下获取负的回报。另外，在当前的试行编号是t的情况下，在步骤S230中获取的回报r是r_{t+1}。

[0220] 在本实施方式中，也谋求式(2)所示的行为价值函数Q的更新，但为了使行为价值函数Q不断适当地更新而必须将示出行为价值函数Q的多层神经网络最佳化(使θ最佳化)。然后，为了通过图8所示的多层神经网络使行为价值函数Q适当地输出而需要成为该输出的目标的教师数据。即，当以使多层神经网络的输出与目标的误差最小化的方式改善θ时，可期待使多层神经网络最佳化。

[0221] 但是，在本实施方式中，在学习未完成的阶段没有行为价值函数Q的认知，确定目标是困难的。因此，在本实施方式中，通过式(2)的第二项、使所谓的TD误差最小化的目标函数来实施示出多层神经网络的θ的改善。即，将$(r_{t+1}+\gamma \max_{a'}Q(s_{t+1},a';\theta_t))$作为目标，以使目标与$Q(s_t,a_t;\theta_t)$的误差最小化的方式学习θ。但是，由于目标$(r_{t+1}+\gamma \max_{a'}Q(s_{t+1},a';\theta_t))$包含有学习对象的θ，因此在本实施方式中经过某一程度的试行次数来固定目标(例如，通过最后学习到的θ(初次学习时是θ的初始值)来固定)。在本实施方式中，预先决定固定目标的试行次数即既定次数。

[0222] 由于在这样的前提下进行学习，因此当在步骤S230中评价回报时，学习部41b算出目标函数(步骤S235)。即，学习部41b算出用于评价试行的各自中的TD误差的目标函数(例如，与TD误差的2次方的期望值成比例的函数、TD误差的2次方的总和等)。另外，由于TD误差在固定有目标的状态下被算出，因此当将被固定后的目标标记为$(r_{t+1}+\gamma \max_{a'}Q(s_{t+1},a';\theta_-))$时，TD误差是$(r_{t+1}+\gamma \max_{a'}Q(s_{t+1},a';\theta_-)-Q(s_t,a_t;\theta_t))$。在该TD误差的式中回报$r_{t+1}$是根据行为$a_t$在步骤S230中得到的回报。

[0223] 此外，$\max_{a'}Q(s_{t+1},a';\theta_-)$是将根据行为$a_t$在步骤S225中算出的状态$s_{t+1}$作为通过被固定后的$\theta_-$确定的多层神经网络的输入的情况下得到的输出之中的最大值。$Q(s_t,a_t;\theta_t)$是在将选择了行为a_t的前一状态s_t作为通过试行编号t的阶段的θ_t确定的多层神经网络的输入的情况下得到的输出之中、对应于行为a_t的输出的值。

[0224] 当算出目标函数时，学习部41b判定学习是否已结束(步骤S240)。在本实施方式中，预先决定用于判定TD误差是否充分小的阈值，在目标函数在阈值以下的情况下，学习部41b判定为学习已结束。

[0225] 在步骤S240中在未判定为学习已结束的情况下，学习部41b更新行为价值(步骤S245)。即，学习部41b基于TD误差的θ的偏微分来确定用于使目标函数减小的θ的变化，并使θ变化。当然，在此，能够通过各种方法使θ变化，例如，能够采用RMSProp等的梯度下降法。此外，也可以适当实施学习率等的调整。根据以上的处理，能够以使行为价值函数Q接近目标的方式使θ变化。

[0226] 但是，在本实施方式中，如上述那样，由于目标被固定，因此学习部41b进一步地进行是否更新目标的判定。具体地，学习部41b判定是否进行了既定次数的试行(步骤S250)，在步骤S250中，在判定为进行了既定次数的试行的情况下，学习部41b更新目标(步骤S255)。即，学习部41b将在算出目标时参照的θ更新为最新的θ。这之后，学习部41b反复进行步骤S215之后的处理。另一方面，在步骤S250中，若未判定为进行了既定次数的试行的情况下，则学习部41b跳过步骤S255反复进行步骤S215之后的处理。

[0227] 在步骤S240中在判定为学习已结束的情况下，学习部41b更新学习信息44e(步骤S260)。即，学习部41b将通过学习得到的θ作为在机器人3的作业时应该参照的θ记录在学习信息44e。在记录有包含该θ的学习信息44e的情况下，如步骤S110～S130那样在进行机器人3的作业时，控制部43基于参数44a控制机器人3。然后，在该作业的过程中，反复进行基于状态观测部41a的当前的状态的观测和基于学习部41b的行为的选择。当然，这时，学习部41b选择在将状态作为输入而被算出来的输出Q(s,a)之中赋予最大值的行为a。然后，在选择了行为a的情况下，以使成为相当于进行了行为a的状态的值的方式更新参数44a。

[0228] 根据以上的结构，控制部43能够一边选择使行为价值函数Q最大化的行为a一边执行作业。反复进行了多数试行的结果，该行为价值函数Q通过上述的处理而被最佳化。然后，该试行通过算出部41自动进行，能够容易地执行不能够人为地实施的程度的多数试行。因此，根据本实施方式，能够以比人为地决定的动作参数高的概率提高机器人3的作业的品质。

[0229] 进一步地，在本实施方式中，根据行为使作为参数44a的伺服增益变化。因此，能够自动地调整通过人为的调整进行适当的设定较困难的、用于控制马达的伺服增益。进一步地，在本实施方式中，根据行为使作为参数44a的加减速特性变化。因此，能够自动地调整通过人为的调整进行适当的设定较困难的加减速特性。

[0230] 进一步地，在本实施方式中，根据行为而机器人的动作的起始点以及终点不变化。因此，在本实施方式中，能够防止自机器人3所预定的起始点以及终点偏离并进行利用者不打算进行的动作。进一步地，在本实施方式中，根据行为而相对于机器人的示教位置即起始点以及终点不变化。因此，在本实施方式中，能够防止自机器人3所示教的位置偏离并进行利用者不打算进行的动作。另外，在本实施方式中，示教位置是起始点以及终点，但也可以使其他位置为示教位置。例如，在起始点与终点之间在存在应该通过的位置、应该采取的姿势的情况下，它们也可以是示教位置(示教姿势)。

[0231] 进一步地，在本实施方式中，由于基于机器人3所进行的作业的好坏与否来评价行为的回报，因此能够以使机器人3的作业成功的方式使参数最佳化。进一步地，在本实施方式中，由于在作业的所需时间比基准短的情况下将回报评价为正，因此能够容易地算出在短的时间使机器人3作业的动作参数。进一步地，在本实施方式中，由于在机器人3的位置与目标位置的偏离量比基准小的情况下将回报评价为正，因此能够容易地算出使机器人3向目标位置正确地移动的动作参数。

[0232] 进一步地，在本实施方式中，由于在振动强度比基准小的情况下将回报评价为正，因此能够容易地算出使基于机器人3的动作的振动产生的可能性较低的动作参数。进一步地，在本实施方式中，由于在机器人3的位置的超调比基准小的情况下将回报评价为正，因此能够容易地算出机器人3超调的可能性较低的动作参数。进一步地，在本实施方式中，由

于在发生声音比基准小的情况下将回报评价为正，因此能够容易地算出使机器人3产生异常的可能性较低的动作参数。

[0233] 进一步地，根据本实施方式，由于自动使行为价值函数Q最佳化，因此能够容易地算出进行高性能的动作的动作参数。此外，由于自动地进行行为价值函数Q的最佳化，因此也能够自动地进行最佳的动作参数的算出。

[0234] 进一步地，在本实施方式中，由于在机器人3中通过通用使用的力觉传感器P来获取机器人3的位置，因此能够基于在机器人3中通用使用的传感器来算出位置信息。

[0235] 进一步地，在本实施方式中，学习部41b对作为状态变量的机器人3的动作结果进行实际测量，使动作参数最佳化。因此，能够在通过机器人3进行了作业的实际环境下配合地使动作参数最佳化。因此，能够以使成为对应于机器人3的使用环境的动作参数的方式进行最佳化。

[0236] 进一步地，在本实施方式中，状态观测部41a在使作为末端执行器的夹钳23设置于机器人3的状态下观测状态变量。此外，学习部41b在使作为末端执行器的夹钳23设置于机器人3的状态下执行作为行为的参数44a的变更。根据该结构，能够容易地算出适合于进行使用了作为末端执行器的夹钳23的动作的机器人3的动作参数。

[0237] 进一步地，在本实施方式中，状态观测部41a在作为末端执行器的夹钳23把持有对象物的状态下观测状态变量。此外，学习部41b在作为末端执行器的夹钳23把持有对象物的状态下执行作为行为的参数44a的变更。根据该结构，能够容易地算出适合于通过作为末端执行器的夹钳23把持对象物并进行动作的机器人3的动作参数。

[0238] (4-5)力控制参数的学习：

[0239] 在力控制参数的学习中，也能够选择学习对象的参数，在此，说明其一例。图11是通过与图7同样的模式对力控制参数的学习例进行了说明的图。本例也基于式(2)使行为价值函数Q(s,a)最佳化。因此，视为使最佳化后的行为价值函数Q(s,a)最大化的行为a是最佳的行为，视为示出该行为a的参数44a是被最佳化后的参数。

[0240] 在力控制参数的学习中，使力控制参数变化也相当于行为的决定，使示出学习对象的参数和可以采取的行为的行为信息44d预先记录于存储部44。即，作为学习对象记述于该行为信息44d的力控制参数成为学习对象。在图11中，机器人3中的力控制参数之中的阻抗参数、力控制坐标系、目标力和机器人3的动作的起始点以及终点是学习对象。另外，力控制中的动作的起始点以及终点是示教位置，通过力控制参数的学习而可以变动。此外，力控制坐标系的原点是自机器人3的TCP(工具中心点)的偏移点，是在学习前目标力进行作用的作用点。因此，当力控制坐标系(原点坐标和轴旋转角)和目标力变化时，自TCP的偏移点的位置变化，而可能产生目标力的作用点不是力控制坐标系的原点的情况。

[0241] 关于相对于机器人坐标系的各轴的平移和旋转而定义力控制参数之中的阻抗参数m、k、d。因此，在本实施方式中，能够针对每一轴使三个阻抗参数m、d、k的各个增加或减少，关于增加可以选择十八个行为、关于减少也可以选择十八个行为，共计可以选择三十六个行为(行为a1～a36)。

[0242] 另一方面，力控制坐标系通过以机器人坐标系为基准来表现该坐标系的原点坐标和力控制坐标系的轴的旋转角度而被定义。因此，在本实施方式中，能够进行原点坐标的朝三轴方向的增减和三轴的轴旋转角的增减，关于原点坐标的增加能够进行三个、关于减少

能够进行三个、关于轴旋转角的增减能够进行三个、关于减少能够进行三个行为，共计可以选择十二个行为(行为a37～a48)。目标力由目标力向量来表现，并通过目标力的作用点和力控制坐标系的六轴各自的成分的大小(三轴平移力、三轴扭矩)来定义。因此，在本实施方式中，关于目标力的作用点的朝三轴方向的增减能够进行六个、关于六轴各自的成分的增加能够进行六个、关于减少能够进行六个行为，共计可以选择十八个行为(行为a49～a66)。

[0243] 机器人3的动作的起始点以及终点沿着机器人坐标系的各轴方向能够进行坐标的增减，关于起始点的增减可以选择六个、关于终点的增减可以选择六个的共计可以选择十二个行为(行为a67～a78)。在本实施方式中，使与通过以上那样预先定义的行为的选择项对应的参数作为学习对象记述于行为信息44d。此外，将用于确定各行为的信息(行为的ID、各行为中的增减量等)记述于行为信息44d。

[0244] 在图11所示的例子中，基于机器人3所进行的作业的好坏与否评价回报。即，学习部41b在使力控制参数作为行为a变化了之后，通过该力控制参数使机器人3动作，并执行拾取通过检测部42检测到的对象物的作业。进一步地，学习部41b观测作业的好坏与否，评价作业的好坏与否。然后，学习部41b根据作业的好坏与否决定行为a、状态s、s'的回报。

[0245] 另外，作业的好坏与否不仅包含作业的成功与否(拾取的成功与否等)还包含作业的品质。具体地，学习部41b基于未图示的计时电路获取从作业的开始至结束(从步骤S110的开始至在步骤S125中判定为结束)的所需时间。然后，学习部41b在作业的所需时间比基准短的情况下赋予正(例如+1)的回报，在作业的所需时间比基准长的情况下赋予负(例如-1)的回报。另外，基准可以根据各种要素来确定，例如，可以是上次的作业的所需时间，可以是过去的最短所需时间，并且也可以是预先决定的时间。

[0246] 进一步地，学习部41b在作业的各工序中基于U1转换机器人3的编码器E1～E6的输出并获取夹钳23的位置。然后，学习部41b在经过进行整定以前的预定期间获取在作业的各工序中获取到的夹钳23的位置，并获取该期间中的振动强度。然后，学习部41b在该振动强度的程度比基准小的情况下赋予正的、在比基准大的情况下赋予负的回报。另外，基准可以根据各种要素来确定，例如，可以是上次的振动强度的程度，可以是过去的最小的振动强度的程度，并且也可以是预先决定的振动强度的程度。

[0247] 振动强度的程度可以通过各种方法来确定，能够采用各种方法：自目标位置的背离的积分值、阈值以上的振动产生的期间等。另外，预定期间能够设为各种期间，若是从工序的起始点到终点的期间，则评价基于动作中的振动强度的回报，若使工序的尾期设为预定期间，则评价基于残留振动的强度的回报。另外，在力控制中，多数情况是基于前者的振动强度的回报的一方较重要。若基于前者的振动强度的回报的一方较重要，则也可以构成为不评价基于后者的残留振动的强度的回报。

[0248] 进一步地，学习部41b在经过进行整定以前的预定期间获取在作业的各工序的尾期获取到的夹钳23的位置，并获取自该期间中的目标位置的背离的最大值作为超调量。然后，学习部41b在该超调量比基准小的情况下赋予正的回报，在比基准大的情况下赋予负的回报。另外，基准可以根据各种要素来确定，例如，可以是上次的超调量的程度，可以是过去的最小的超调量，并且也可以是预先决定的超调量。

[0249] 进一步地，在本实施方式中，在控制装置40、机器人1～3、作业台等的至少一处安装有声音采集装置，学习部41b获取示出在作业中声音采集装置获取到的声音的信息。然

后，学习部41b在作业中的发生声音的大小比基准小的情况下赋予正的回报，在比基准大的情况下赋予负的回报。另外，基准可以根据各种要素来确定，例如，可以是上次的作业或工序的发生声音的大小的程度，可以是过去的发生声音的大小的最小值，并且也可以是预先决定的大小。此外，发生声音的大小可以通过声压的最大值来评价，也可以通过预定期间内的声压的统计值(平均值等)来评价，能够采用各种结构。

[0250] 另外，在力控制参数的学习中，在动作参数的学习中设为了回报的、自目标位置的背离不包含于回报中。即，在力控制参数的学习中，由于工序的起始点、终点可能根据学习而变动，因此不包含于回报中。

[0251] 在进行了作为行为a的参数的变化之后使机器人3动作，状态观测部41a观测状态，从而能够确定在当前的状态s下在采用了行为a的情况下的下一状态s'。另外，在基于机器人1、2的对象物的检测完成后，关于机器人3执行本例的力控制参数的学习。

[0252] 在图11所示的例子中，在状态变量中包含有马达M1～M6的电流、编码器E1～E6的值、力觉传感器P的输出。因此，状态观测部41a观测向马达M1～M6供给的电流值作为伺服43d的控制结果。该电流值相当于由马达M1～M6输出的扭矩。此外，使编码器E1～E6的输出基于对应关系U1转换为机器人坐标系中的TCP的位置。因此，状态观测部41a观测机器人3具备的夹钳23的位置信息。

[0253] 在本实施方式中，将在机器人的运动中通过力觉传感器P检测到的输出积分，从而能够算出机器人的位置。即，状态观测部41a基于对应关系U2在机器人坐标系中将朝运动中的TCP的作用力进行积分，从而获取TCP的位置。因此，在本实施方式中状态观测部41a也利用力觉传感器P的输出观测机器人3具备的夹钳23的位置信息。另外，状态可以通过各种方法来观测，也可以使不进行上述的转换的值(电流值、编码器、力觉传感器的输出值)作为状态来观测。

[0254] 状态观测部41a不是直接观测行为即阻抗参数、力控制坐标系、工序的起始点以及终点的调整结果，而是将调整的结果、将由机器人3得到的变化作为马达M1～M6的电流、编码器E1～E6的值、力觉传感器P的输出来观测。因此，变为间接观测行为的影响，在该意义上，本实施方式的状态变量是可以进行从力控制参数的变化直接进行推定较困难的变化的状态变量。

[0255] 此外，马达M1～M6的电流、编码器E1～E6的值、力觉传感器P的输出直接示出了机器人1～3的动作，该动作直接示出了作业的好坏与否。因此，作为状态变量，观测马达M1～M6的电流、编码器E1～E6的值、力觉传感器P的输出，从而能够进行人为地改善较困难的参数的改善，并以有效地提高作业的品质的方式使力控制参数最佳化。其结果，能够以比人为地决定的力控制参数高的概率来算出进行高性能的动作的力控制参数。

[0256] (4-6)力控制参数的学习例：

[0257] 接下来，说明力控制参数的学习例。使示出在学习的过程中参照的变量、函数的信息作为学习信息44e存储于存储部44。即，算出部41采用了如下结构：通过反复进行状态变量的观测、对应于该状态变量的行为的决定和由该行为得到的回报的评价而使行为价值函数Q(s,a)收敛。因此，在本例中，使在学习的过程中状态变量、行为和回报的时序的值依次不断记录于学习信息44e。

[0258] 另外，在本实施方式中，通过力控制模式来执行力控制参数的学习(在仅进行位置

控制的位置控制模式中不进行力控制参数的学习)。为了执行力控制模式中的学习,可以使仅通过力控制模式构成的作业作为机器人3的机器人程序44b而生成,在使包含任意模式的作业作为机器人3的机器人程序44b而生成的状况下,也可以仅使用其中的力控制模式进行学习。

[0259] 行为价值函数Q(s,a)可以通过各种方法来算出,也可以基于多次试行来算出,在此,对通过DQN使行为价值函数Q最佳化的例子进行说明。使利用于行为价值函数Q的最佳化的多层神经网络在上述的图8中示意性地示出。若是观测图11所示的那样的状态的本例,则由于机器人3中的马达M1～M6的电流、编码器E1～E6的值、力觉传感器P的输出(六轴的输出)为状态,因此状态s的数量M=18。此外,若是可以选择图11所示的行为的本例,则由于能够选择七十八个行为,因此N=78。当然,行为a的内容、数量(N的值)、状态s的内容、数量(M的值)也可以根据试行编号t而变化。

[0260] 在本实施方式中,也将用于确定该多层神经网络的参数(为了从输入得到输出而需要的信息)作为学习信息44e记录于存储部44。在此,也将在学习的过程中可能变化的多层神经网络的参数标记为θ。当使用该θ时,也能够将上述的行为价值函数$Q(s_t,a_{1t})$～$Q(s_t,a_{Nt})$标记为$Q(s_t,a_{1t};\theta_t)$～$Q(s_t,a_{Nt};\theta_t)$。

[0261] 接下来,沿着图9所示的流程图来说明学习处理的顺序。力控制参数的学习处理可以在机器人3的运用过程中实施,也可以在实际运用之前事先执行学习处理。在此,按照在实际运用之前事先执行学习处理的结构(当使示出多层神经网络的θ最佳化时,是保存该信息并在下一次之后的运用中利用的结构)说明学习处理。

[0262] 当开始学习处理时,算出部41将学习信息44e初始化(步骤S200)。即,算出部41确定在开始学习时参照的θ的初始值。初始值可以通过各种方法来决定,在过去未进行学习的情况下,可以使任意值、随机值等为θ的初始值,也可以准备模拟机器人3、对象物的模拟环境并基于该环境将所学习或所推定的θ作为初始值。

[0263] 在过去进行了学习的情况下,将该学习完毕的θ作为初始值采用。此外,在过去进行了关于类似的对象的学习的情况下,也可以使该学习中的θ设为初始值。过去的学习可以是用户使用机器人3来进行,也可以是机器人3的制造商在机器人3的销售前进行。在该情况下,也可以构成为制造商根据对象物、作业的种类预备好多个初始值的组,在用户学习时选择初始值。当决定θ的初始值时,使该初始值作为当前的θ的值存储于学习信息44e。

[0264] 接下来,算出部41将参数初始化(步骤S205)。在此,由于力控制参数是学习对象,因此算出部41将力控制参数初始化。即,若是未进行学习的状态,则算出部41将包含于通过示教而生成的参数44a的力控制参数作为初始值来设定。若是过去进行了某种学习的状态,则算出部41将包含于在学习时最后利用的参数44a的力控制参数作为初始值来设定。

[0265] 接下来,状态观测部41a观测状态变量(步骤S210)。即,控制部43参照参数44a以及机器人程序44b控制机器人3(相当于上述的步骤S110～S130)。这之后,状态观测部41a观测向马达M1～M6供给的电流值。此外,状态观测部41a获取编码器E1～E6的输出,基于对应关系U1转换为机器人坐标系中的TCP的位置。进一步地,状态观测部41a将力觉传感器P的输出积分,获取TCP的位置。

[0266] 接下来,学习部41b算出行为价值(步骤S215)。即,学习部41b参照学习信息44e获取θ,向示出学习信息44e的多层神经网络输入最新的状态变量,算出N个行为价值函数Q

$(s_t,a_{1t};\theta_t)\sim Q(s_t,a_{Nt};\theta_t)$。

[0267] 另外，最新的状态变量在初次的执行时是步骤S210的观测结果，在第二次之后的执行时是步骤S225的观测结果。此外，试行编号t在初次的执行时为零，在第二次之后的执行时为1以上的值。在过去未实施学习处理的情况下，由于未使学习信息44e示出的θ最佳化，因此作为行为值函数Q的值可能变为不正确的值，但通过步骤S215以后的处理的反复，从而使行为值函数Q逐渐不断最佳化。此外，在步骤S215以后的处理的反复中，能够使状态s、行为a、回报r与各试行编号t建立对应地存储于存储部44，在任意定时参照。

[0268] 接下来，学习部41b选择行为并执行(步骤S220)。在本实施方式中，进行视为使行为价值函数Q(s,a)最大化的行为a是最佳的行为的处理。因此，学习部41b确定在步骤S215中算出来的N个行为价值函数$Q(s_t,a_{1t};\theta_t)\sim Q(s_t,a_{Nt};\theta_t)$的值之中最大的值。然后，学习部41b选择赋予了最大的值的行为。例如，若在N个行为价值函数$Q(s_t,a_{1t};\theta_t)\sim Q(s_t,a_{Nt};\theta_t)$之中$Q(s_t,a_{1t};\theta_t)$是最大值，则学习部41b选择行为$a_{Nt}$。

[0269] 当选择行为时，学习部41b使与该行为对应的参数44a变化。例如，在图11所示的例子中，在选择了使与机器人坐标系中的x轴相关的阻抗参数m增加一定值的行为a1的情况下，学习部41b在力控制参数示出的与x轴相关的阻抗参数m增加一定值。当进行参数44a的变化时，控制部43参照该参数44a控制机器人3，使执行一系列的作业。另外，在本实施方式中，每当进行行为选择则执行一系列的作业，但也可以构成为每当进行行为选择则执行一系列的作业的一部分(构成为执行构成一系列的作业的多个工序的至少一工序)。

[0270] 接下来，状态观测部41a观测状态变量(步骤S225)。即，状态观测部41a进行与步骤S210中的状态变量的观测同样的处理，作为状态变量而获取向马达M1～M6供给的电流值、基于编码器E1～E6的输出确定的TCP的位置、基于力觉传感器P的输出确定的TCP的位置。另外，在当前的试行编号是t的情况下(所选择的行为是a_t的情况下)，在步骤S225中获取的状态s是s_{t+1}。

[0271] 接下来，学习部41b评价回报(步骤S230)。即，学习部41b基于未图示的计时电路获取从作业的开始至结束的所需时间，在作业的所需时间比基准短的情况下获取正的回报，在作业的所需时间比基准长的情况下获取负的回报。进一步地，学习部41b获取作业的各工序中的夹钳23的位置，基于在作业的各工序中获取到的夹钳23的位置获取振动强度。然后，在该振动强度比基准小的情况下获取正的回报，在比基准大的情况下获取负的回报。在使一系列的作业由多个工序构成的情况下，可以获取各工序的回报之和，也可以获取统计值(平均值等)。

[0272] 进一步地，学习部41b基于在作业的各工序的尾期获取到的夹钳23的位置来获取超调量。然后，学习部41b在使在该超调量比基准小的情况下获取正的回报、在比基准大的情况下获取负的回报的一系列的作业由多个工序构成的情况下，可以获取各工序的回报之和，也可以获取统计值(平均值等)。

[0273] 进一步地，学习部41b获取示出在作业中声音采集装置获取到的声音的信息。然后，学习部41b在作业中的发生声音的大小比基准小的情况下获取正的回报，在比基准大的情况下获取负的回报。另外，在当前的试行编号是t的情况下，在步骤S230中获取的回报r是r_{t+1}。

[0274] 在本实施方式中，也谋求式(2)所示的行为价值函数Q的更新，但为了使行为价值

函数Q不断适当地更新而必须将示出行为价值函数Q的多层神经网络最佳化(使θ最佳化)。然后,为了通过图8所示的多层神经网络使行为价值函数Q适当地输出而需要成为该输出的目标的教师数据。即,当以使多层神经网络的输出与目标的误差最小化的方式改善θ时,可期待使多层神经网络最佳化。

[0275] 但是,在本实施方式中,在学习未完成的阶段没有行为价值函数Q的认知,确定目标是困难的。因此,在本实施方式中,通过式(2)的第二项、使所谓的TD误差最小化的目标函数来实施示出多层神经网络的θ的改善。即,将$(r_{t+1}+\gamma \max_{a'}Q(s_{t+1},a';\theta_t))$作为目标,以使目标与$Q(s_t,a_t;\theta_t)$的误差最小化的方式学习θ。但是,由于目标$(r_{t+1}+\gamma \max_{a'}Q(s_{t+1},a';\theta_t))$包含有学习对象的θ,因此在本实施方式中经过某一程度的试行次数来固定目标(例如,通过最后学习到的θ(初次学习时是θ的初始值)来固定)。在本实施方式中,预先决定固定目标的试行次数即既定次数。

[0276] 由于在这样的前提下进行学习,因此当在步骤S230中评价回报时,学习部41b算出目标函数(步骤S235)。即,学习部41b算出用于评价试行的各自中的TD误差的目标函数(例如,与TD误差的2次方的期望值成比例的函数、TD误差的2次方的总和等)。另外,由于TD误差在固定有目标的状态下被算出,因此当将被固定后的目标标记为$(r_{t+1}+\gamma \max_{a'}Q(s_{t+1},a';\theta^-))$时,TD误差是$(r_{t+1}+\gamma \max_{a'}Q(s_{t+1},a';\theta^-)-Q(s_t,a_t;\theta_t))$。在该TD误差的式中回报$r_{t+1}$是根据行为$a_t$在步骤S230中得到的回报。

[0277] 此外,$\max_{a'}Q(s_{t+1},a';\theta^-)$是将根据行为$a_t$在步骤S225中算出的状态$s_{t+1}$作为通过被固定后的$\theta^-$确定的多层神经网络的输入的情况下得到的输出之中的最大值。$Q(s_t,a_t;\theta_t)$是在将选择了行为a_t的前一状态s_t作为通过试行编号t的阶段的θ_t确定的多层神经网络的输入的情况下得到的输出之中、对应于行为a_t的输出的值。

[0278] 当算出目标函数时,学习部41b判定学习是否已结束(步骤S240)。在本实施方式中,预先决定用于判定TD误差是否充分小的阈值,在目标函数在阈值以下的情况下,学习部41b判定为学习已结束。

[0279] 在步骤S240中在未判定为学习已结束的情况下,学习部41b更新行为价值(步骤S245)。即,学习部41b基于TD误差的θ的偏微分来确定用于使目标函数减小的θ的变化,并使θ变化。当然,在此,能够通过各种方法使θ变化,例如,能够采用RMSProp等的梯度下降法。此外,也可以适当实施学习率等的调整。根据以上的处理,能够以使行为价值函数Q接近目标的方式使θ变化。

[0280] 但是,在本实施方式中,如上述那样,由于目标被固定,因此学习部41b进一步地进行是否更新目标的判定。具体地,学习部41b判定是否进行了既定次数的试行(步骤S250),在步骤S250中,在判定为进行了既定次数的试行的情况下,学习部41b更新目标(步骤S255)。即,学习部41b将在算出目标时参照的θ更新为最新的θ。这之后,学习部41b反复进行步骤S215之后的处理。另一方面,在步骤S250中,若未判定为进行了既定次数的试行的情况下,则学习部41b跳过步骤S255反复进行步骤S215之后的处理。

[0281] 在步骤S240中在判定为学习已结束的情况下,学习部41b更新学习信息44e(步骤S260)。即,学习部41b将通过学习得到的θ作为在机器人3的作业时应该参照的θ记录在学习信息44e。在记录有包含该θ的学习信息44e的情况下,如步骤S110～S130那样在进行机器人3的作业时,控制部43基于参数44a控制机器人3。然后,在该作业的过程中,反复进行基于状

态观测部41a的当前的状态的观测和基于学习部41b的行为的选择。当然，这时，学习部41b选择在将状态作为输入而被算出来的输出Q(s,a)之中赋予最大值的行为a。然后，在选择了行为a的情况下，以使成为相当于进行了行为a的状态的值的方式更新参数44a。

[0282] 根据以上的结构，控制部43能够一边选择使行为价值函数Q最大化的行为a一边执行作业。反复进行了多数试行的结果，该行为价值函数Q通过上述的处理而被最佳化。然后，该试行通过算出部41自动进行，能够容易地执行不能够人为地实施的多数试行。因此，根据本实施方式，能够以比人为地决定的力控制参数高的概率提高机器人3的作业的品质。

[0283] 进一步地，在本实施方式中，根据行为使作为参数44a的阻抗参数变化。因此，能够自动地调整通过人为的调整进行适当的设定较困难的阻抗参数。进一步地，在本实施方式中，根据行为使作为参数44a的起始点和终点变化。因此，以能够高性能地进行力控制的方式自动地调整人为地设定了的起始点、终点。

[0284] 进一步地，在本实施方式中，根据行为使作为参数44a的力控制坐标系变化。其结果，使自机器人3的TCP的偏移点的位置变化。因此，能够自动地调整通过人为的调整进行适当的设定较困难的、自TCP的偏移点的位置。进一步地，在本实施方式中，根据行为可以使作为参数44a的目标力变化。因此，能够自动地调整通过人为的调整进行适当的设定较困难的目标力。特别是，由于人为地使力控制坐标系和目标力的组合理想化是困难的，因此使这些群组自动地调整的结构是有用的。

[0285] 进一步地，在本实施方式中，由于基于机器人3所进行的作业的好坏与否来评价行为的回报，因此能够以使机器人3的作业成功的方式使参数最佳化。进一步地，在本实施方式中，由于在作业的所需时间比基准短的情况下将回报评价为正，因此能够容易地算出在短的时间使机器人3作业的动作参数。

[0286] 进一步地，在本实施方式中，由于在振动强度比基准小的情况下将回报评价为正，因此能够容易地算出使机器人3的动作产生振动的可能性较低的力控制参数。进一步地，在本实施方式中，由于在机器人3的位置的超调比基准小的情况下将回报评价为正，因此能够容易地算出机器人3超调的可能性较低的力控制参数。进一步地，在本实施方式中，由于在发生声音比基准小的情况下将回报评价为正，因此能够容易地算出使机器人3产生异常的可能性较低的力控制参数。

[0287] 进一步地，根据本实施方式，由于自动使行为价值函数Q最佳化，因此能够容易地算出进行高性能的力控制的力控制参数。此外，由于自动地进行行为价值函数Q的最佳化，因此也能够自动地进行最佳的力控制参数的算出。

[0288] 进一步地，在本实施方式中，由于在机器人3中通过通用使用的力觉传感器P来获取机器人3的位置，因此能够基于在机器人3中通用使用的传感器来算出位置信息。

[0289] 进一步地，在本实施方式中，学习部41b对作为状态变量的机器人3的动作结果进行实际测量，使力控制参数最佳化。因此，能够在通过机器人3进行了作业的实际环境下配合地使力控制参数最佳化。因此，能够以使成为对应于机器人3的使用环境的力控制参数的方式进行最佳化。

[0290] 进一步地，在本实施方式中，状态观测部41a在使作为末端执行器的夹钳23设置于机器人3的状态下观测状态变量。此外，学习部41b在使作为末端执行器的夹钳23设置于机器人3的状态下执行作为行为的参数44a的变更。根据该结构，能够容易地算出适合于进行

使用了作为末端执行器的夹钳23的动作的机器人3的力控制参数。

[0291] 进一步地，在本实施方式中，状态观测部41a在作为末端执行器的夹钳23把持有对象物的状态下观测状态变量。此外，学习部41b在作为末端执行器的夹钳23把持有对象物的状态下执行作为行为的参数44a的变更。根据该结构，能够容易地算出适合于通过作为末端执行器的夹钳23把持对象物并进行动作的机器人3的力控制参数。

[0292] (5)其他实施方式：

[0293] 以上的实施方式是用于实施本发明的一例，此外也能够采用各种实施方式。例如，控制装置可以内置于机器人，也可以设于与机器人的设置场所不同的场所、例如外部的服务器等。此外，控制装置可以由多个装置构成，也可以由与控制部43和算出部41不同的装置构成。此外，控制装置可以是与机器人控制器、示教盒、PC、网络连接的服务器等，也可以包含有它们。进一步地，可以省略上述的实施方式的一部分的结构，也可以使处理的顺序变动或省略。进一步地，在上述的实施方式中，关于TCP设定了目标位置、目标力的初始向量，但也可以在其他位置、例如关于力觉传感器P的传感器坐标系的原点、关于螺丝的末端等设定目标位置、目标力的初始向量。

[0294] 机器人只要能够通过任意形态的可动部实施任意作业即可。末端执行器是利用于与对象物的作业相关的部位，也可以安装有任意工具。对象物只要是成为机器人的作业对象的物体即可，可以是由末端执行器把持的物体，也可以是由末端执行器具备的工具处置的物体，各种物体可以成为对象物。

[0295] 作用于机器人的目标力只要是在通过力控制驱动该机器人时作用于机器人的目标力即可，例如，将通过力觉传感器等的力检测部检测的力(或从该力算出的力)控制为特定的力时，该力成为目标力。此外，可以以使由力觉传感器以外的传感器、例如加速度传感器检测的力(或从该力算出的力)成为目标力的方式进行控制，也可以以使加速度、角速度成为特定的值的方式进行控制。

[0296] 进一步地，在上述的学习处理中，每当进行试行则通过θ的更新来更新行为价值，并在进行至既定次数的试行后固定目标，但也可以在进行多次的试行之后进行θ的更新。例如，可列举如下结构：在进行至第一既定次数的试行后使目标固定，在进行至第二既定次数(＜第一既定次数)的试行后固定θ。在该情况下，构成为在第二既定次数的试行后基于第二既定次数份的样本来更新θ，进一步地在试行次数超过第一既定次数的情况下由最新的θ更新目标。

[0297] 进一步地，在学习处理中，可以采用众所周知的各种方法，例如，也可以进行体验再现、回报的Clipping等。进一步地，在图8中，存在P个(P是1以上的整数)层DL，在各层中存在多个节点，但各层的构造能够采用各种构造。例如，层的数量、节点的数量能够采用各种数量，作为激活函数也能够采用各种函数，并且也可以使网络构造成为卷积神经网络构造等。此外，输入、输出的形态也不限定于图8所示的例子，例如，也可以采用至少利用输入状态s和行为a的结构、使将行为价值函数Q最大化的行为a作为独热(one-hot)向量输出的结构的例子。

[0298] 在上述的实施方式中，一边基于行为价值函数通过贪婪(greedy)策略进行行为一边使行为价值函数最佳化，从而视为相对于最佳化后的行为价值函数的greedy策略是最佳策略。该处理是所谓的价值反复法，但也可以通过其他方法、例如策略反复法进行学习。进

一步地，在状态s、行为a、回报r等的各种变量中，也可以进行各种正规化。

[0299] 作为机械学习的方法，采用各种方法，也可以通过基于行为价值函数Q的ε-greedy策略进行试行。此外，作为强化学习的方法也并不限定于上述那样的Q学习，也可以使用SARSA等的方法。此外，也可以利用使策略的模式与行为价值函数的模式分别地模式化的方法、例如Actor-Critic算法。若利用Actor-Critic算法，则可以构成为定义示出策略的actor即μ(s;θ)和示出行为价值函数的critic即Q(s,a;θ)，按照对μ(s;θ)施加有噪音的策略生成行为并进行试行，基于试行结果更新actor和critic，从而学习策略和行为价值函数。

[0300] 算出部只要能够使用机械学习算出学习对象的参数即可，作为参数，只要是光学参数、图像处理参数、动作参数、力控制参数的至少一个即可。机械学习只要是使用样本参数学习更好的参数的处理即可，除了上述的强化学习以外，也能够采用通过有教师学习、聚类等各种方法来学习各参数的结构。

[0301] 光学系统能够拍摄对象物。即，具备获取使包含对象物的区域在视野内的图像的结构。作为光学系统的结构要素，如上述那样，优选包含拍摄部、照明部，此外也可以包含有各种结构要素。此外，如上述那样，可以使拍摄部、照明部能够通过机器人的臂而移动，可以通过二维的移动结构而能够移动，并且也可以是固定的。当然，拍摄部、照明部也可以是能够更换的。此外，在光学系统中使用的光(基于拍摄部的检测光、照明部的输出光)的带域并不限定于可见光带域，能够采用使用红外线、紫外线、X射线等的任意电磁波的结构。

[0302] 光学参数只要是可以使光学系统的状态变化的值即可，使在由拍摄部、照明部等构成的光学系统中用于直接或间接确定状态的数值等成为光学参数。例如，不仅示出拍摄部、照明部等的位置、角度等的值，而且示出拍摄部、照明部的种类的数值(ID、型号等)也可以成为光学参数。

[0303] 检测部能够基于在根据算出的光学参数的光学系统中的拍摄结果检测对象物。即，检测部具备如下结构：通过学习到的光学参数使光学系统动作，拍摄对象物，并基于拍摄结果执行对象物的检测处理。

[0304] 检测部只要能够检测对象物即可，如上述的实施方式那样，除了检测对象物的位置姿势的结构外，还可以构成为检测对象物的有无，能够采用各种结构。另外，对象物的位置姿势例如能够通过基于三轴中的位置和相对于三轴的旋转角的六个参数来定义，但当然根据需要也可以不考虑任意数量的参数。例如，只要是设置于平面上的对象物，则假设与至少一个位置相关的参数为已知，也可以从对象物中除去。此外，只要是以固定的朝向设置于平面的对象物，则也可以从对象物中将与姿势相关的参数除去。

[0305] 对象物只要是成为由光学系统拍摄、检测的对象的物体即可，能够设想成为机器人的作业对象的工件、工件的周边的物体、机器人的一部分等各种物体。此外，作为基于检测结果检测对象物的方法也能够采用各种方法，可以通过图像的特征量提取来检测对象物，也可以通过对象物的动作(人等的可动物体等的检测)来检测对象物，可以采用各种方法。

[0306] 控制部能够基于对象物的检测结果来控制机器人。即，控制部具备根据对象物的检测结果来决定机器人的控制内容的结构。因此，机器人的控制除了上述那样的用于抓住对象物的控制外还可以进行各种控制。例如，可设想基于对象物进行机器人的定位的控制、

基于对象物使机器人的动作开始或结束的控制等各种控制。

[0307] 机器人的形态可以是各种形态，除了上述的实施方式那样的垂直多关节机器人以外还可以是正交机器人、水平多关节机器人、双臂机器人等。此外，也可以使各种形态的机器人组合。当然，轴的数量、臂的数量、末端执行器的形态等能够采用各种形态。例如，使拍摄部21、照明部22安装在存在于机器人3的上方的平面，拍摄部21、照明部22能够在该平面上移动。

[0308] 状态观测部只要能够观测根据行为等的试行而变化了的结果即可，可以通过各种传感器等来观测状态，也可以构成为进行使从某一状态向其他状态变化的控制，若未观测到控制的失败(错误等)则视为观测到该其他状态。基于前者的传感器的观测除了位置等的检测外还包含基于拍摄传感器的图像的获取。

[0309] 进一步地，上述的实施方式中的行为、状态、回报是举例，也可以是包含其他行为、状态、回报的结构、省略了任意行为、状态的结构。例如，在能够更换拍摄部21、照明部22的机器人1、2中，能够将拍摄部21、照明部22的种类的变更作为行为来选择，也能够作为状态来观测种类。也可以基于接触判定部43c的判定结果决定回报。即，在学习部41b中的学习过程中，在接触判定部43c判定为在作业中未设想到的物体与机器人已接触的情况下，能够采用将该之前的行为的回报设定为负的结构。根据该结构，能够以使机器人不与设想外的物体接触的方式使参数44a最佳化。

[0310] 此外，例如，在光学参数的最佳化时，也可以构成为通过机器人1～3进行基于对象物的检测结果的作业(例如，上述的拾取作业等)，学习部41b基于对象物的检测结果并基于机器人1～3所进行的作业的好坏与否来评价行为的回报。该结构例如可列举如下结构：在图7所示的回报中，代替对象物的检测或除了对象物的检测，还将作业的成功与否(例如，拾取的成功与否)设为回报。

[0311] 作业的成功与否例如由能够判定作业的成功与否的工序(拾取的工序等)中的步骤S120的判定结果等来定义。在该情况下，在行为、状态中，也可以包含有与机器人1～3的动作相关的行为、状态。进一步地，在该结构中，优选将由具备拍摄部21以及照明部22的光学系统拍摄到机器人1～3的作业对象即对象物的图像设为状态。根据该结构，能够以使机器人的作业成功的方式使光学参数最佳化。另外，作为为了学习光学参数、动作参数、力控制参数而观测的状态的图像，也可以是由拍摄部21拍摄到的图像本身，也可以是对由拍摄部21拍摄到的图像进行了图像处理(例如，上述的平滑化处理、锐化处理等)之后的图像。

[0312] 进一步地，也可以采用不是使光学参数、动作参数、力控制参数的各个单个地最佳化而是使这些参数之中的两种以上最佳化的结构。例如，在图7所示的例子中，若构成为包含使动作参数、力控制参数变化的行为，则能够与光学参数一起使动作参数、力控制参数最佳化。在该情况下，基于最佳化后的动作参数、力控制参数控制机器人1～3。根据该结构，能够使进行伴随对象物的检测的作业的参数最佳化，能够执行提高对象物的检测精度的学习。

[0313] 图像处理参数只要是可以使作为对象物的拍摄结果的图像变化的值即可，并不限定于图3所示的例子，也可以追加或删除。例如，图像处理的有无、图像处理的强度、图像处理的顺序等、用于确定执行的图像处理的算法的数值(包含示出处理顺序等的图表等)等可以成为图像处理参数。更具体地，作为图像处理，可列举二值化处理、直线检测处理、圆检测

处理、颜色检测处理、OCR处理等。

[0314] 进一步地，图像处理也可以是组合有多个种类的图像处理的处理。例如，也可以组合圆检测处理和OCR处理，进行“识别圆内的文字的处理”这样的处理。无论怎样，示出各图像处理的有无、强度的参数可以成为图像处理参数。此外，这些图像处理参数的变化可以成为行为。

[0315] 动作参数并不限定于上述的实施方式所列举的参数。例如，也可以在成为学习对象的动作参数中包含有用于基于机器人1～3具备的惯性传感器来进行控制的伺服增益。即，在通过基于惯性传感器的输出的控制环路控制马达M1～M6的结构中，也可以构成为使该控制环路中的伺服增益根据行为而变化。例如，在基于安装于机器人1～3的编码器E1～E6来算出机器人1～3的特定的部位的角速度、通过惯性传感器的一种即陀螺传感器来检测该特定的部位的角速度、两者的差分乘以陀螺伺服增益并进行反馈控制的结构中，可列举该陀螺伺服增益根据行为而变化的结构。若是该结构，则能够进行抑制在机器人的特定的部位产生的角速度的振动成分的控制。当然，惯性传感器并不限定于陀螺传感器，在加速度传感器等中在进行同样的反馈控制的结构中也可以构成为加速度增益根据行为而变化。根据以上的结构，能够自动地调整难以通过人为的调整进行适当的设定的、用于基于惯性传感器进行控制的伺服增益。另外，加速度传感器是检测通过机器人的运动产生的加速度的传感器，上述的力觉传感器是检测作用于机器人的力的传感器。通常，加速度传感器和力觉传感器是不同的传感器，但在一方能够代替另一方的功能的情况下，一方也可以作为另一方发挥功能。

[0316] 当然，力控制参数也并不限定于上述的实施方式所列举的参数，并且也可以适当选择成为学习对象的参数。例如，关于目标力，也可以构成为不可以使六轴中的全部成分或一部分的成分作为行为来选择(即是固定的)。该结构在将机器人把持有的对象物插入被固定后的固定对象(细的筒等)的作业中，目标力具有相对固定对象物的某一点而固定的成分，但也能够设想根据机器人的插入作业以使力控制坐标系变化的方式进行学习的结构等。

[0317] 学习部41b也可以构成为在如下的至少一种情况下将回报评价为负：在作业完成前丢下机器人3把持有对象物的情况下，在作业完成前使机器人3的作业对象即对象物的一部分分离的情况下，在机器人3破损了的情况下，在机器人3的作业对象即对象物破损了的情况下。根据在作业完成前丢下机器人3把持的对象物的情况下将回报评价为负的结构，能够容易地算出不使对象物丢下而使作业完成的可能性较高的动作参数、力控制参数。

[0318] 根据在作业完成前使机器人3的作业对象即对象物的一部分分离的情况下将回报评价为负的结构，能够容易地算出不使对象物分离而使作业完成的可能性较高的动作参数、力控制参数。根据在机器人3破损了的情况下将回报评价为负的结构，能够容易地算出不使机器人3破损的可能性较低的动作参数、力控制参数。

[0319] 根据在机器人3的作业对象即对象物破损了的情况下将回报评价为负的结构，能够容易地算出使对象物破损的可能性较低的动作参数、力控制参数。另外，能够采用通过各种传感器、例如拍摄部21等来检测如下情况的结构：在作业完成前是否丢下机器人3把持有的对象物、在作业完成前是否使机器人3的作业对象即对象物的一部分分离、机器人3是否破损了、机器人3的作业对象即对象物是否破损了。

[0320] 进一步地，学习部41b也可以构成为在机器人3的作业正常地完成了的情况下将回报评价为正。根据在机器人3的作业正常地完成了的情况下将回报评价为正的结构，能够容易地算出使机器人3的作业成功的动作参数、力控制参数。

[0321] 进一步地，用于检测机器人3的位置的位置检测部并不限定于上述的实施方式那样的编码器、力觉传感器，也可以是其他传感器、专用的惯性传感器、拍摄部21等的光学传感器、距离传感器等。此外，可以使传感器内置于机器人，也可以配置于机器人的外部。若利用配置于机器人的外部的位置检测部，则能够不影响机器人的动作而算出位置信息。

[0322] 进一步地，算出部41也可构成为基于机器人的不同的多个动作算出多个动作所共用的动作参数、力控制参数。多个动作只要包含有利用被最佳化后的动作参数而执行的动作即可。因此，多个动作可列举不同的种类的多个作业(拾取作业、研磨作业、螺丝紧固作业等)的结构、同种作业(螺丝的大小不同的多个螺丝紧固作业等)的结构等。根据该结构，能够容易地算出能够应用于各种动作的通用的动作参数、力控制参数。

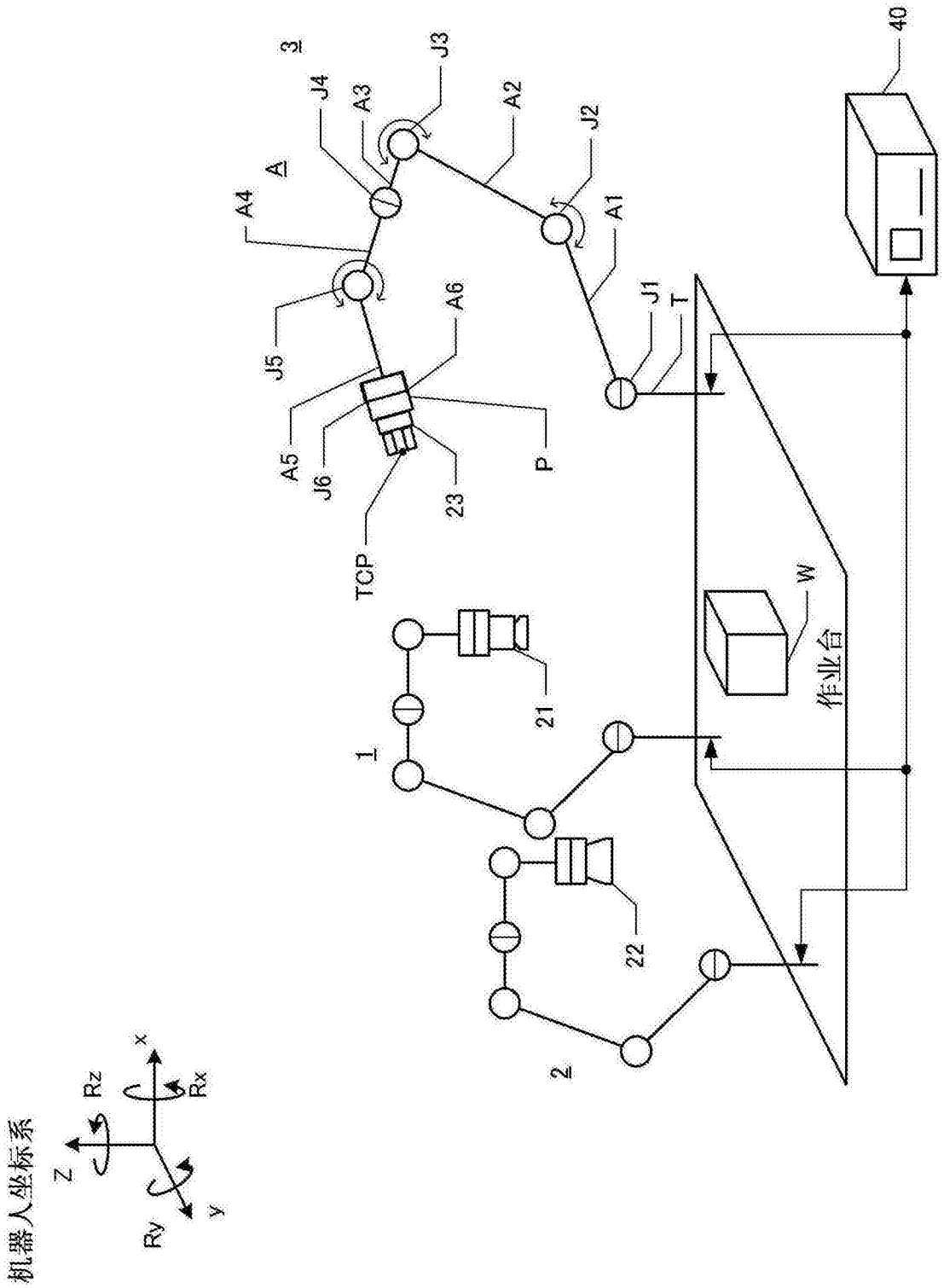

40
3
J3
J4
A3
A2
J2
A
A4
A1
J1
T
J5
A6
A5
J6
23
P
TCP
W
作业台
21
1
22
2
机器人坐标系
x
Rz
Rx
Z
Ry
y

图1

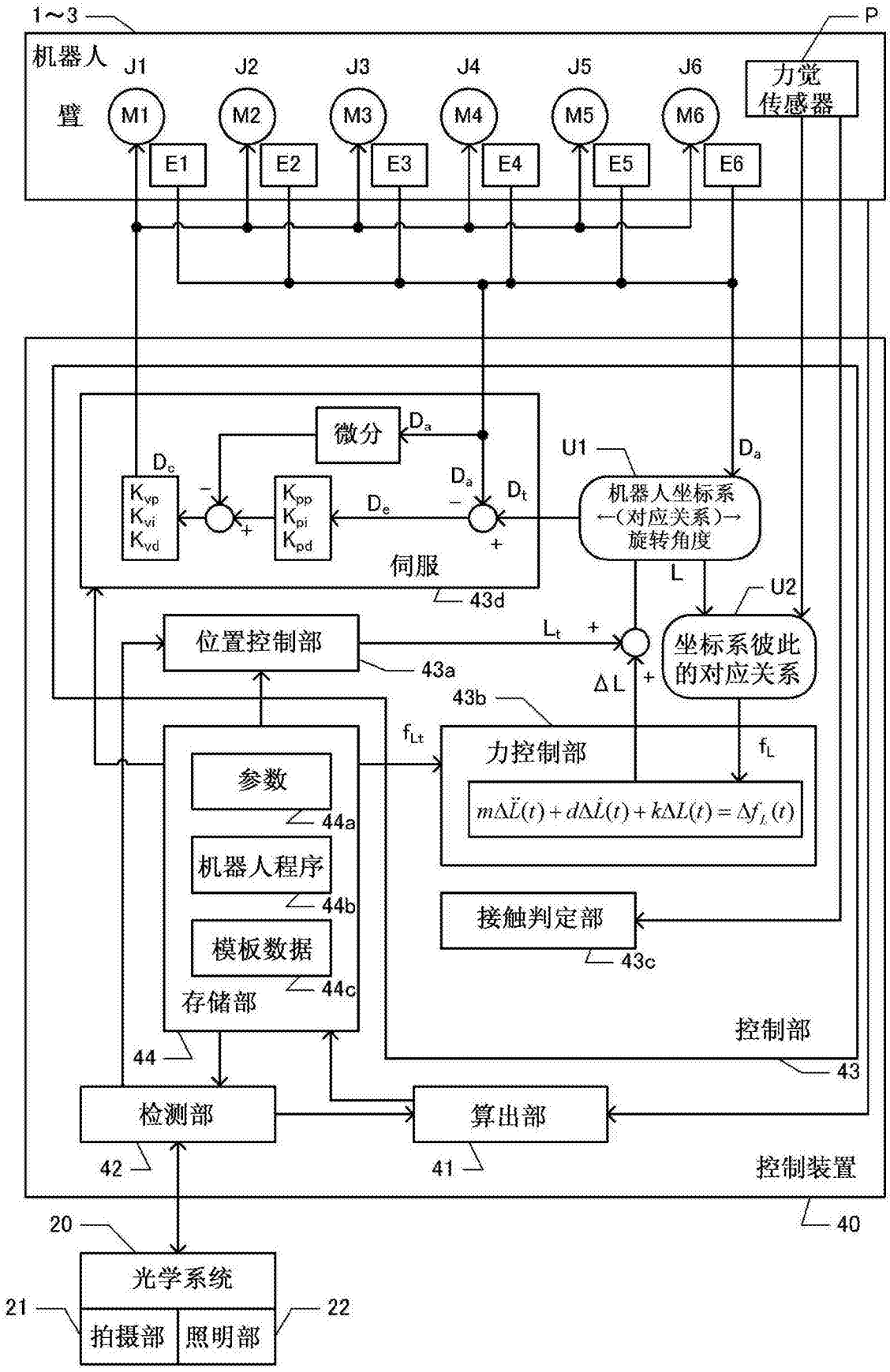

1～3
机器人
臂
J1
J2
J3
J4
J5
J6
M1
M2
M3
M4
M5
M6
E1
E2
E3
E4
E5
E6
P
力觉
传感器
微分
Da
Dc
Kvp
Kvi
Kvd
Kpp
Kpi
Kpd
De
Dt
伺服
43d
U1
机器人坐标系
←(对应关系)→
旋转角度
L
U2
坐标系彼此
的对应关系
位置控制部
43a
Lt
ΔL
43b
fLt
力控制部
fL
参数
44a
机器人程序
44b
模板数据
44c
存储部
44
接触判定部
43c
控制部
43
检测部
42
算出部
41
控制装置
40
20
光学系统
21
拍摄部
照明部
22

图2

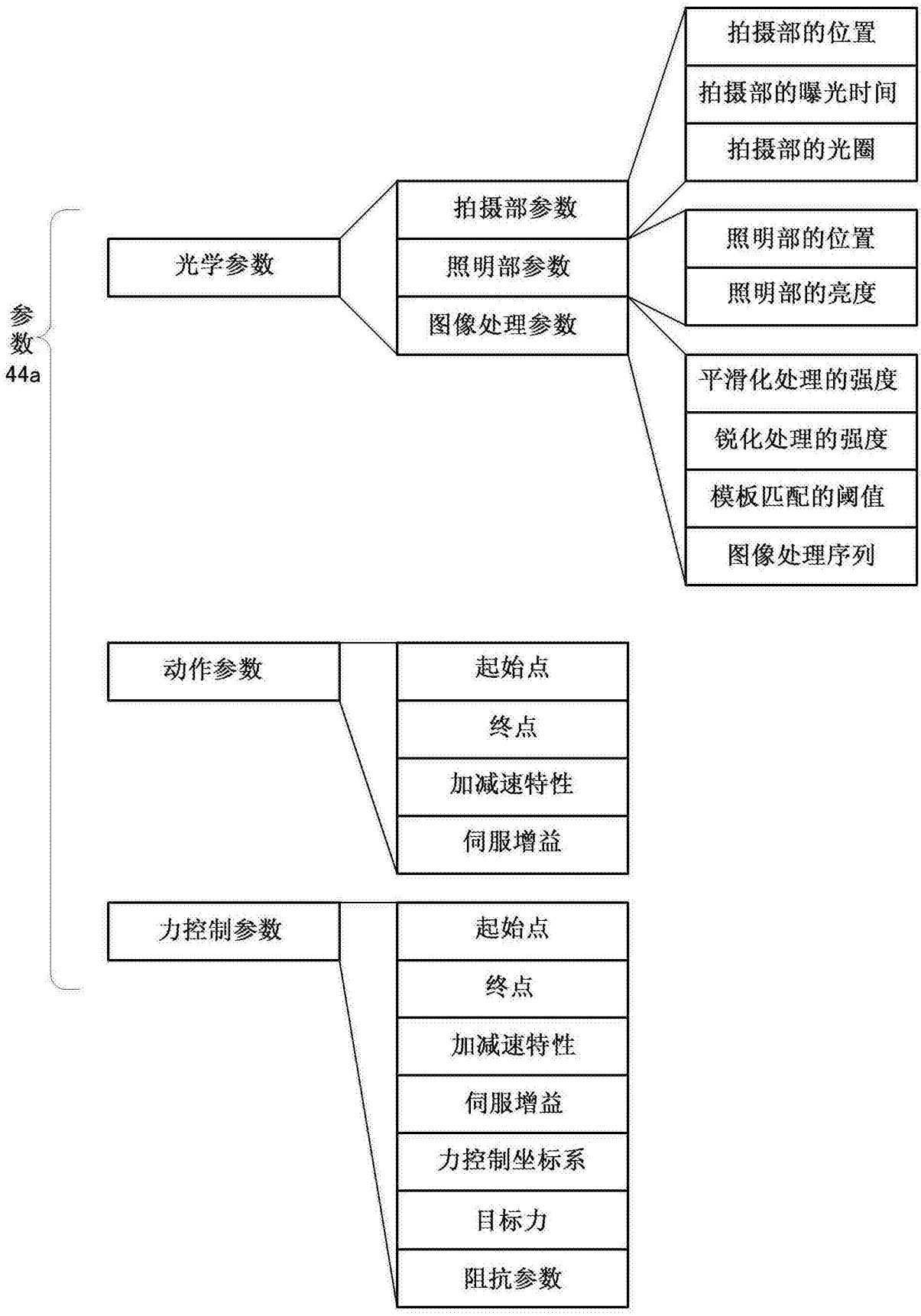

参数44a
光学参数
拍摄部参数
照明部参数
图像处理参数
拍摄部的位置
拍摄部的曝光时间
拍摄部的光圈
照明部的位置
照明部的亮度
平滑化处理的强度
锐化处理的强度
模板匹配的阈值
图像处理序列
动作参数
起始点
终点
加减速特性
伺服增益
力控制参数
起始点
终点
加减速特性
伺服增益
力控制坐标系
目标力
阻抗参数

图3

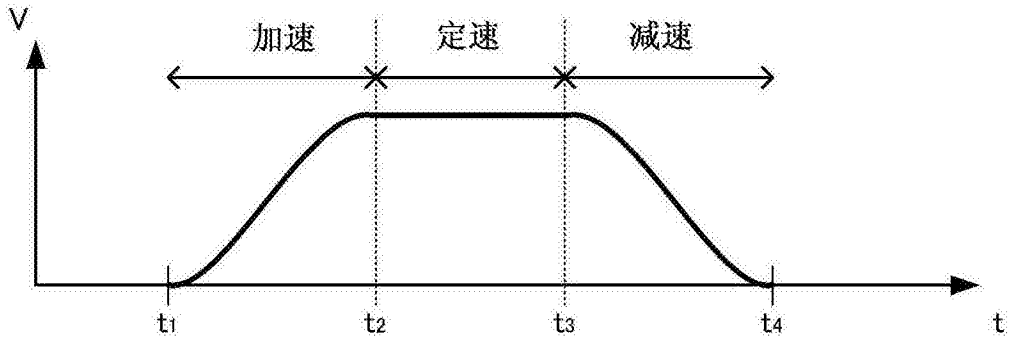

V
加速
定速
减速
t1
t2
t3
t4
t

图4

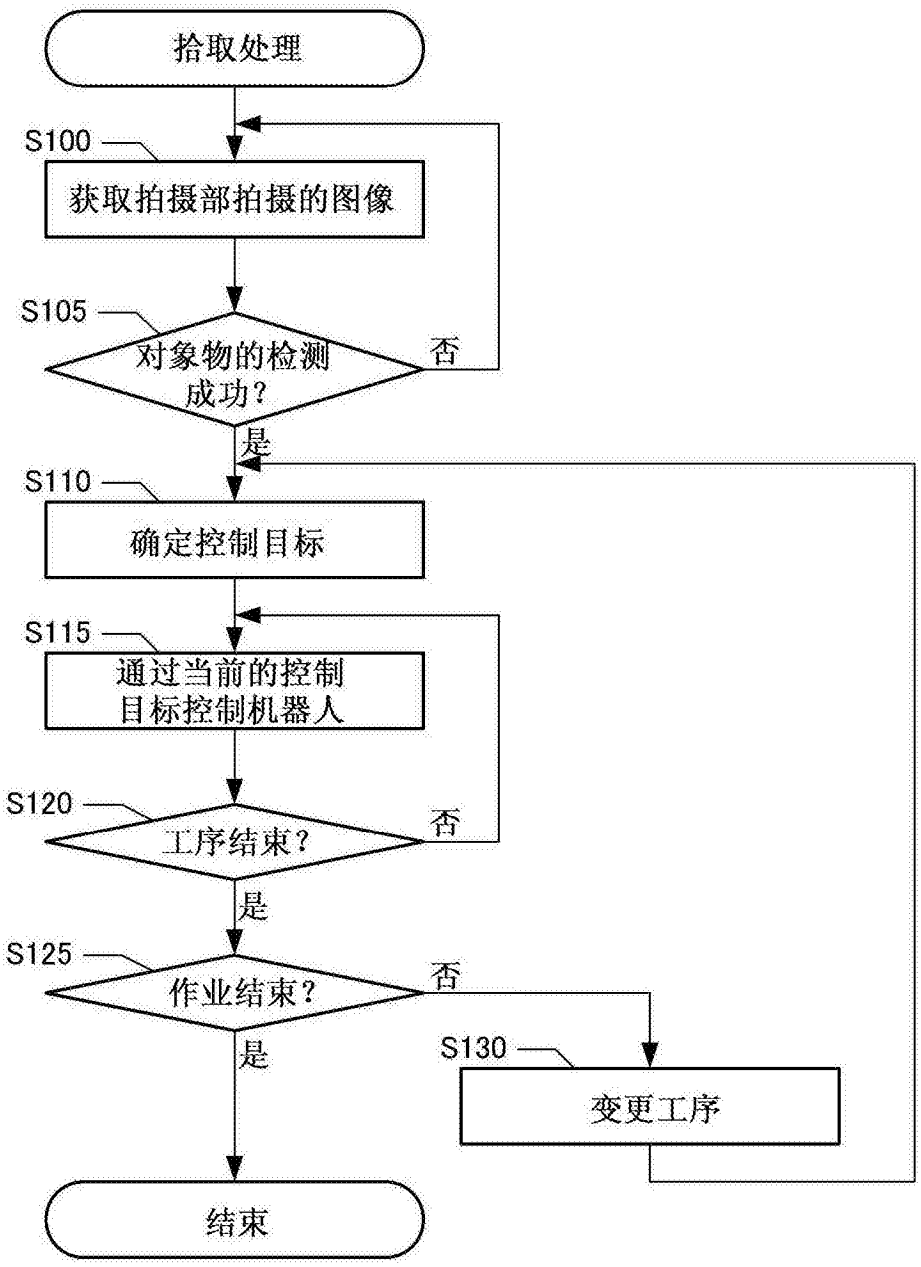

拾取处理
S100
获取拍摄部拍摄的图像
S105
对象物的检测成功？
否
是
S110
确定控制目标
S115
通过当前的控制目标控制机器人
S120
工序结束？
否
是
S125
作业结束？
否
是
S130
变更工序
结束

图5

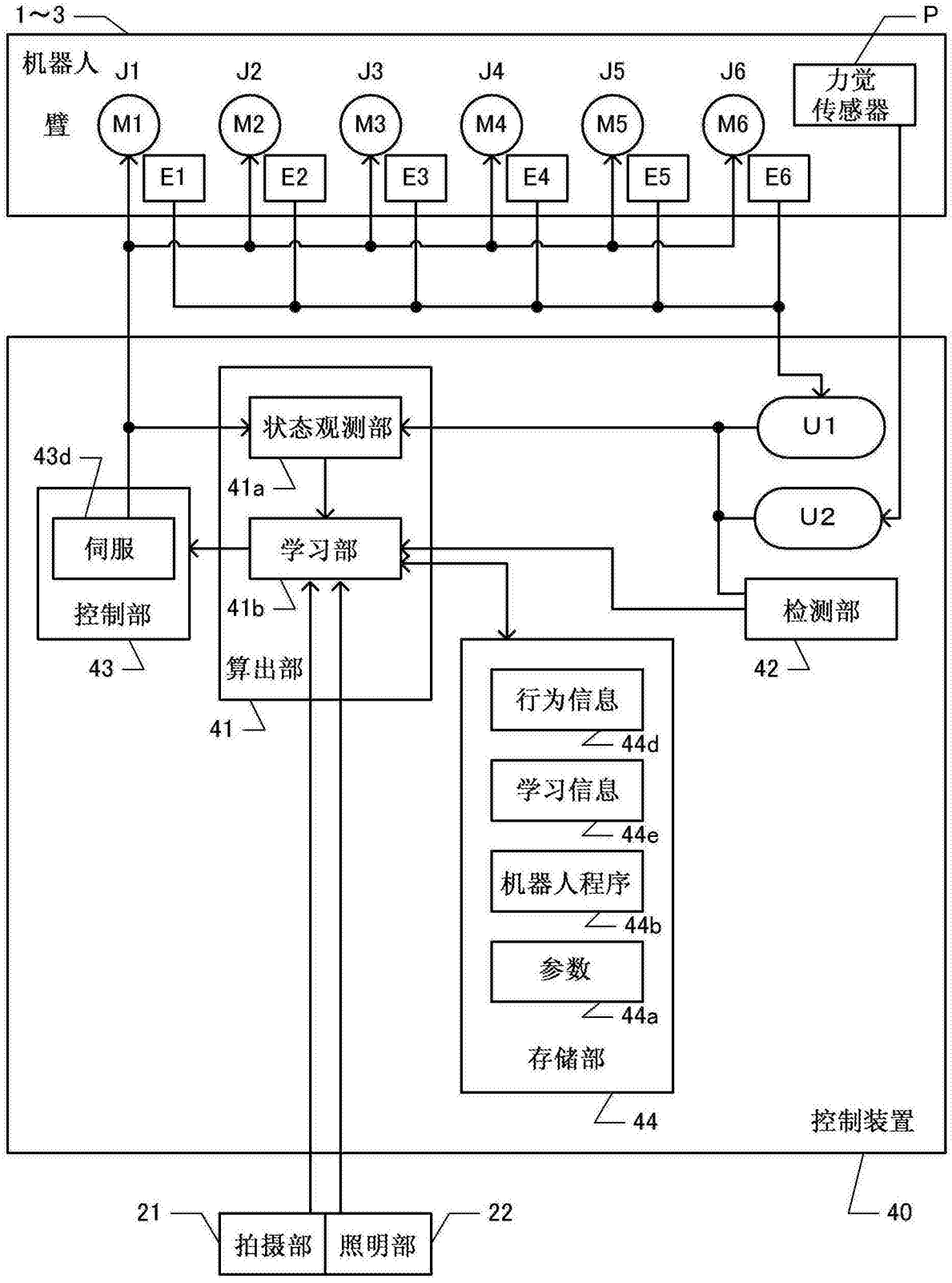

1～3
机器人
臂
J1
J2
J3
J4
J5
J6
M1
M2
M3
M4
M5
M6
E1
E2
E3
E4
E5
E6
P
力觉
传感器
U1
U2
状态观测部
41a
43d
伺服
控制部
43
学习部
41b
算出部
41
检测部
42
行为信息
44d
学习信息
44e
机器人程序
44b
参数
44a
存储部
44
控制装置
40
21
拍摄部
照明部
22

图6

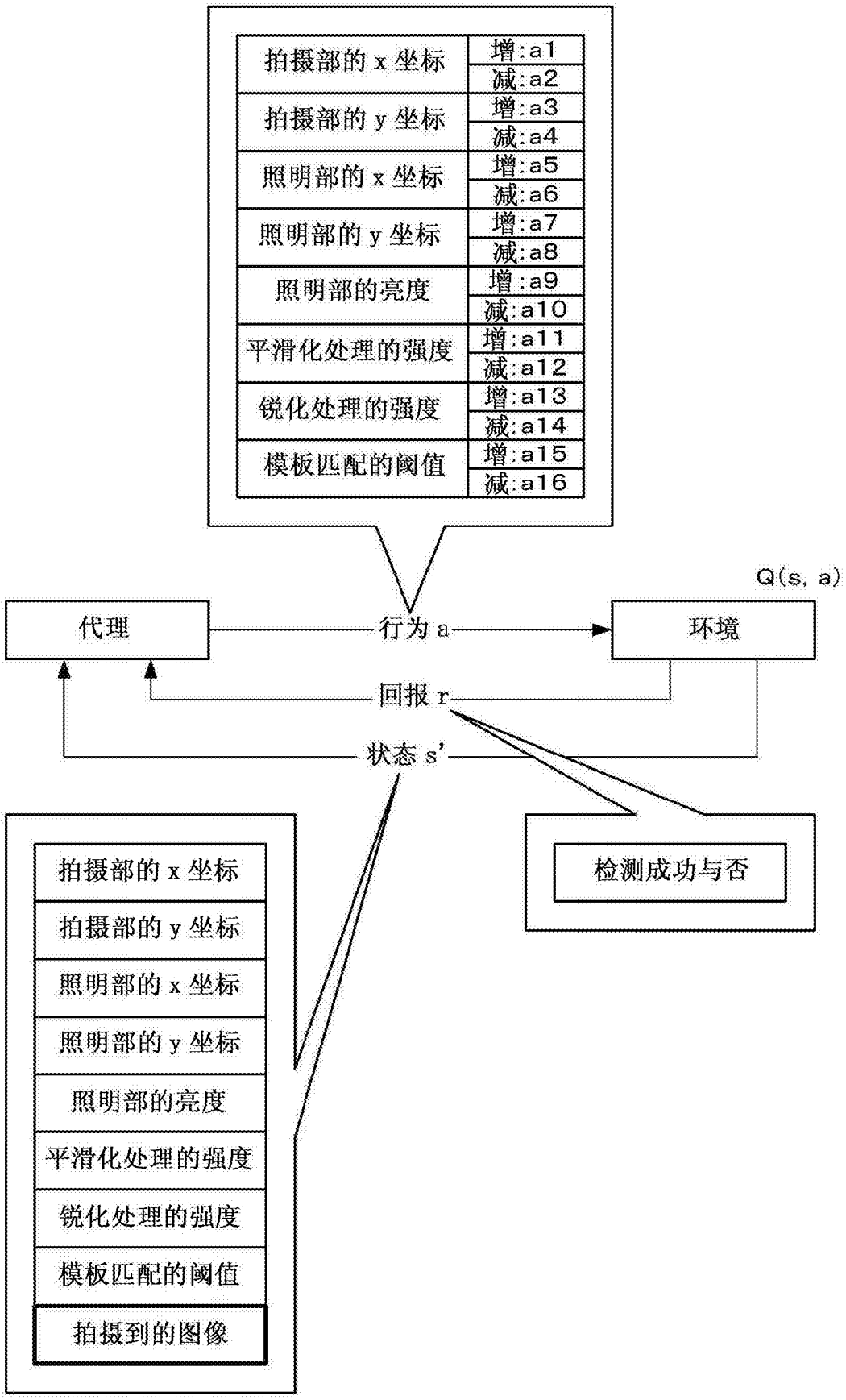

拍摄部的 x 坐标
增:a1
减:a2
拍摄部的 y 坐标
增:a3
减:a4
照明部的 x 坐标
增:a5
减:a6
照明部的 y 坐标
增:a7
减:a8
照明部的亮度
增:a9
减:a10
平滑化处理的强度
增:a11
减:a12
锐化处理的强度
增:a13
减:a14
模板匹配的阈值
增:a15
减:a16
Q(s, a)
代理
行为 a
环境
回报 r
状态 s'
检测成功与否
拍摄部的 x 坐标
拍摄部的 y 坐标
照明部的 x 坐标
照明部的 y 坐标
照明部的亮度
平滑化处理的强度
锐化处理的强度
模板匹配的阈值
拍摄到的图像

图7

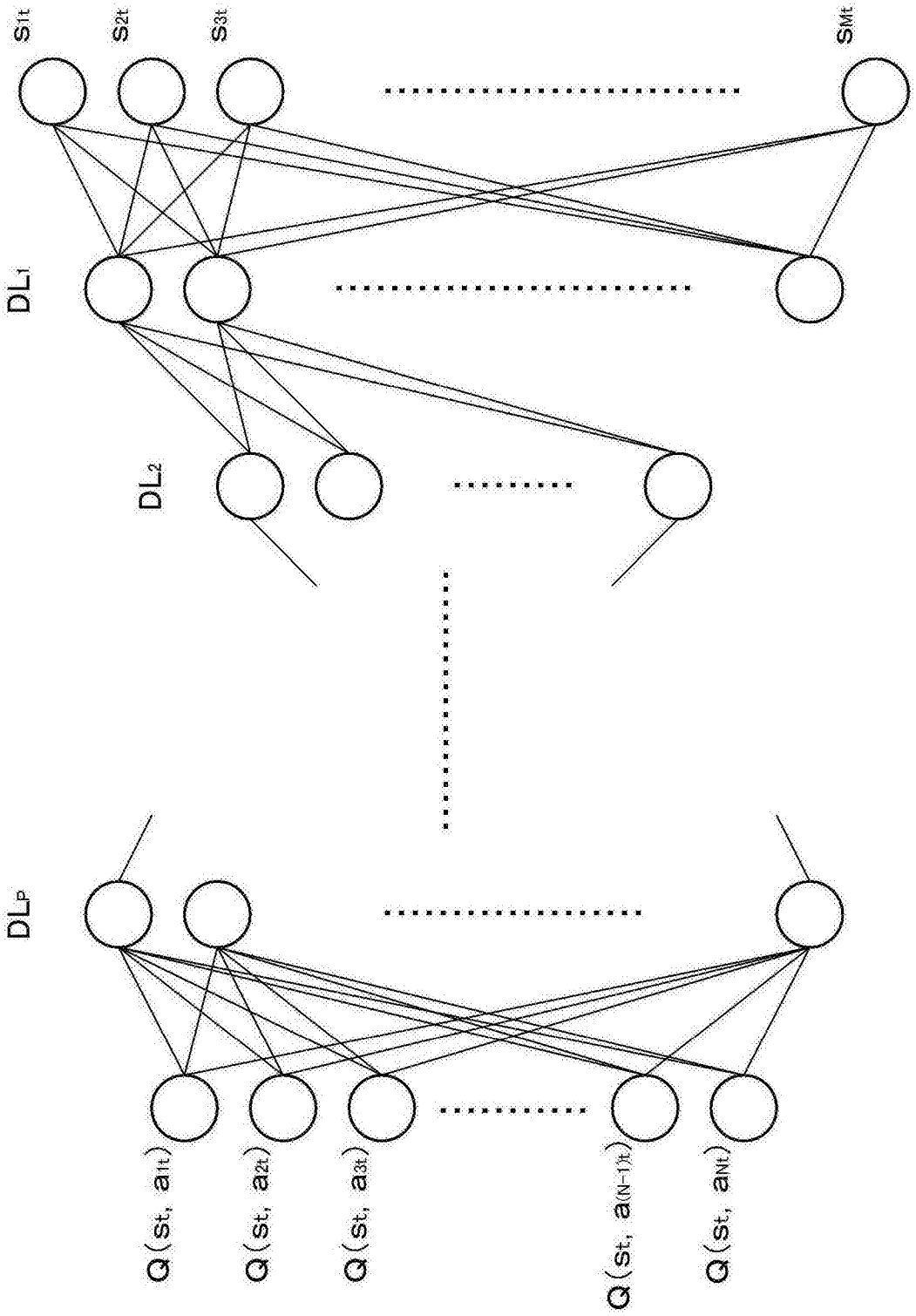

S1t
S2t
S3t
SMt
DL1
DL2
DLP
Q(St, a1t)
Q(St, a2t)
Q(St, a3t)
Q(St, a(N-1)t)
Q(St, aNt)

图8

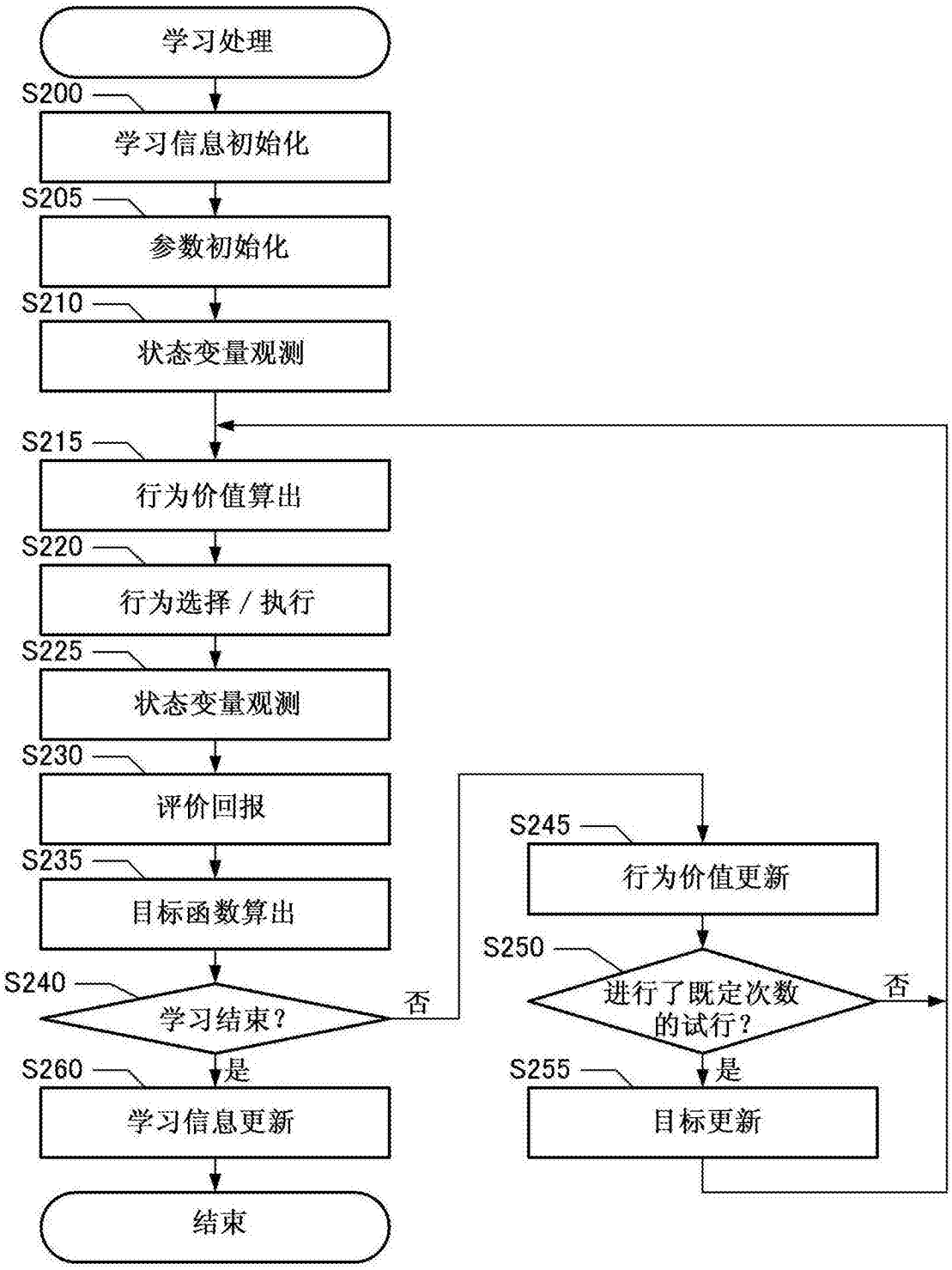

学习处理
S200
学习信息初始化
S205
参数初始化
S210
状态变量观测
S215
行为价值算出
S220
行为选择／执行
S225
状态变量观测
S230
评价回报
S235
目标函数算出
S240
学习结束？
否
是
S260
学习信息更新
结束
S245
行为价值更新
S250
进行了既定次数的试行？
否
是
S255
目标更新

图9

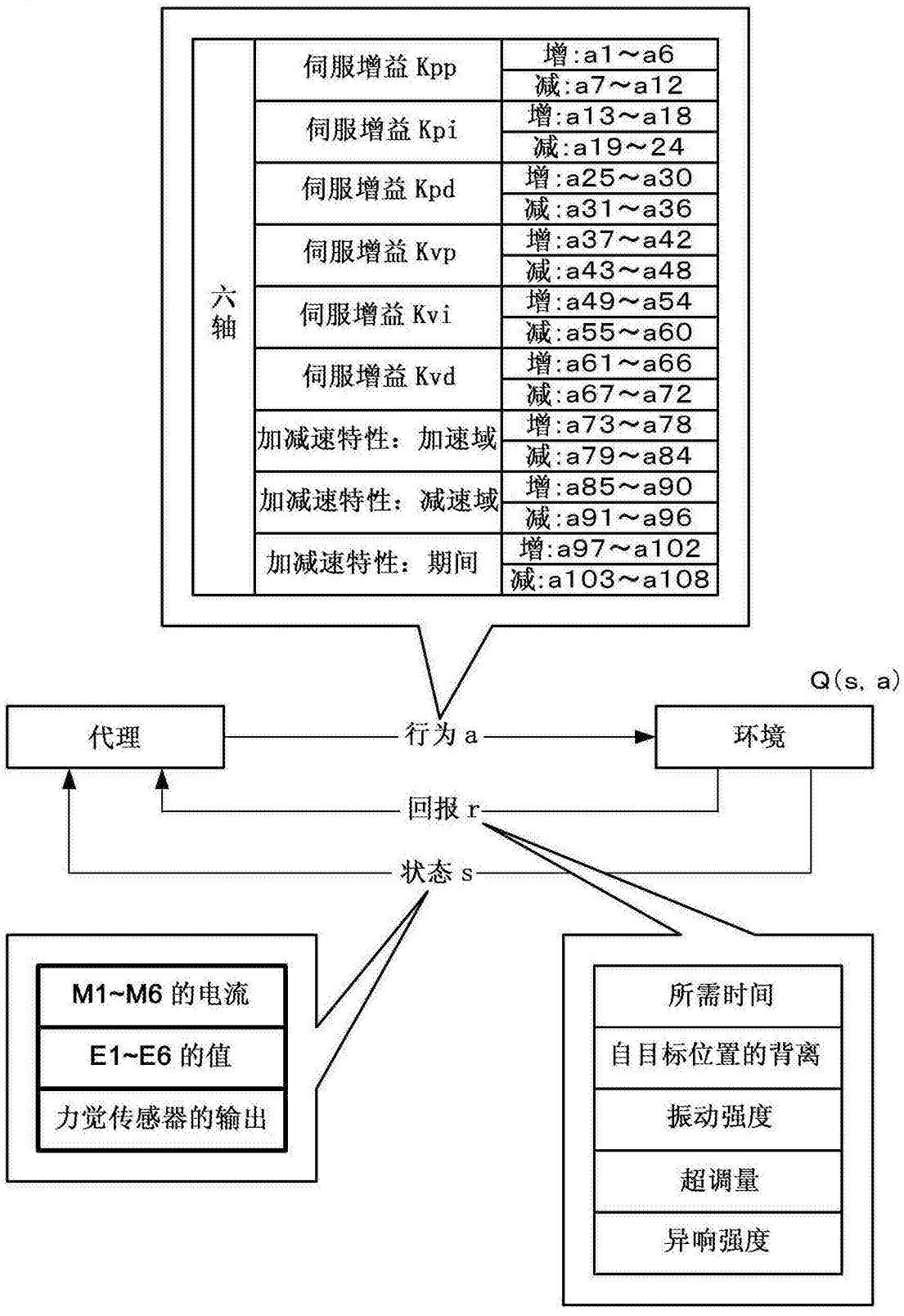

六轴
伺服增益 Kpp
增:a1～a6
减:a7～a12
伺服增益 Kpi
增:a13～a18
减:a19～24
伺服增益 Kpd
增:a25～a30
减:a31～a36
伺服增益 Kvp
增:a37～a42
减:a43～a48
伺服增益 Kvi
增:a49～a54
减:a55～a60
伺服增益 Kvd
增:a61～a66
减:a67～a72
加减速特性：加速域
增:a73～a78
减:a79～a84
加减速特性：减速域
增:a85～a90
减:a91～a96
加减速特性：期间
增:a97～a102
减:a103～a108
Q(s, a)
代理
行为 a
环境
回报 r
状态 s
M1~M6 的电流
E1~E6 的值
力觉传感器的输出
所需时间
自目标位置的背离
振动强度
超调量
异响强度

图10

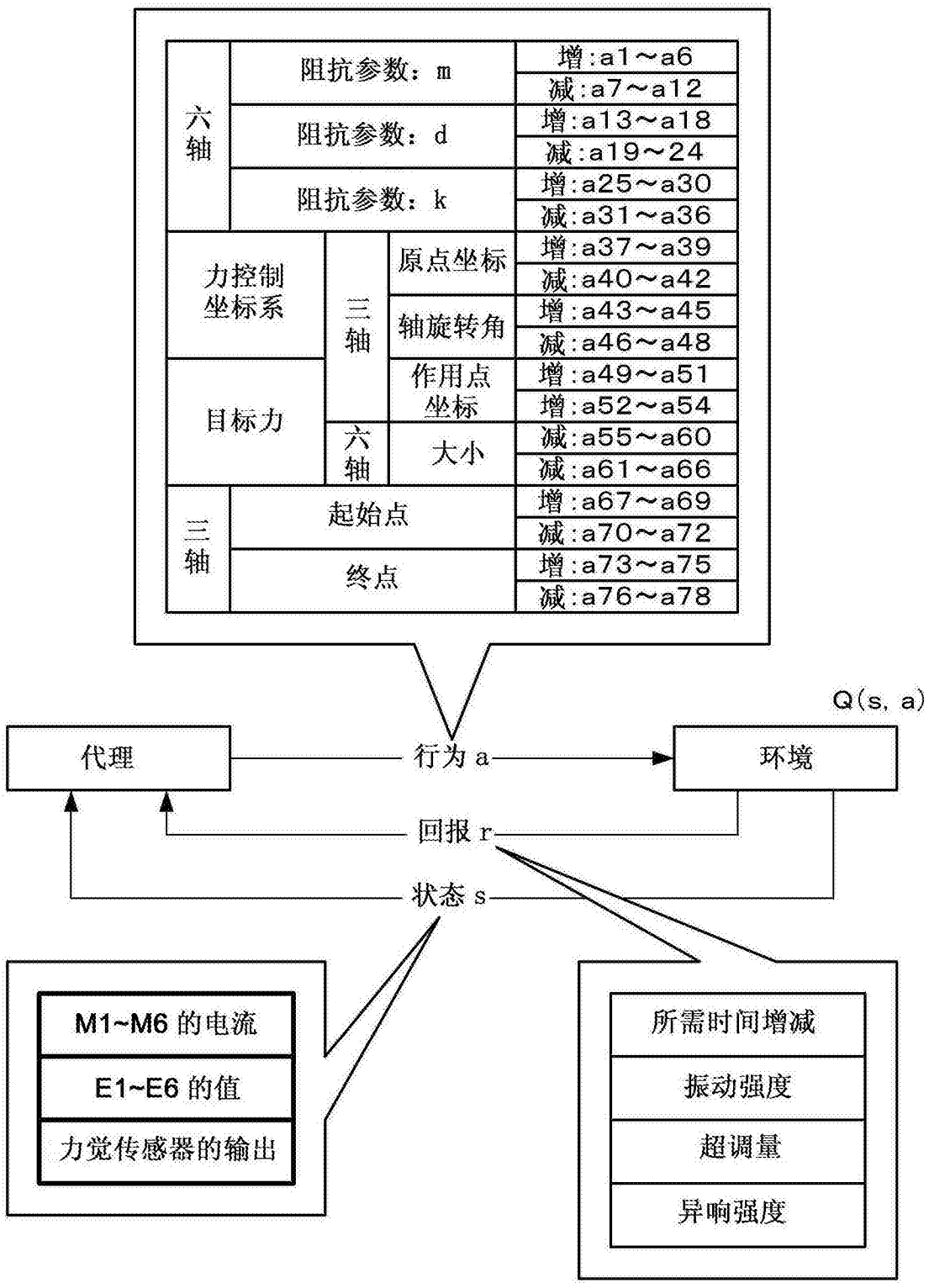

六轴
阻抗参数：m
增:a1～a6
减:a7～a12
阻抗参数：d
增:a13～a18
减:a19～24
阻抗参数：k
增:a25～a30
减:a31～a36
力控制坐标系
三轴
原点坐标
增:a37～a39
减:a40～a42
轴旋转角
增:a43～a45
减:a46～a48
目标力
作用点坐标
增:a49～a51
增:a52～a54
六轴
大小
减:a55～a60
减:a61～a66
三轴
起始点
增:a67～a69
减:a70～a72
终点
增:a73～a75
减:a76～a78
Q(s, a)
代理
行为 a
环境
回报 r
状态 s
M1~M6 的电流
E1~E6 的值
力觉传感器的输出
所需时间增减
振动强度
超调量
异响强度

图11