(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2018-171663
(P2018-171663A)
(43)公開日　平成30年11月8日(2018.11.8)

| (51)Int.Cl. | | | F I | | | テーマコード（参考） |
|---|---|---|---|---|---|---|
| B25J | 13/00 | (2006.01) | B25J | 13/00 | Z | 3C269 |
| G05B | 19/4155 | (2006.01) | G05B | 19/4155 | V | 3C707 |

審査請求 有　請求項の数 8　ＯＬ　（全 18 頁）

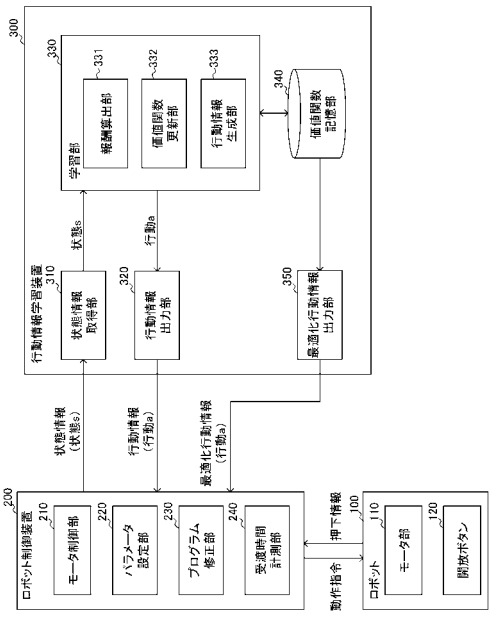
(21)出願番号　特願2017-69866(P2017-69866)
(22)出願日　平成29年3月31日(2017.3.31)

(71)出願人　390008235
ファナック株式会社
山梨県南都留郡忍野村忍草字古馬場３５８０番地
(74)代理人　100106002
弁理士　正林　真之
(74)代理人　100165157
弁理士　芝　哲央
(74)代理人　100160794
弁理士　星野　寛明
(72)発明者　山本　知之
山梨県南都留郡忍野村忍草字古馬場３５８０番地　ファナック株式会社内

最終頁に続く

(54)【発明の名称】行動情報学習装置、ロボット制御システム及び行動情報学習方法

(57)【要約】
【課題】作業者がロボットとの協調作業を行いやすくするための行動情報学習装置、ロボット制御システム及び行動情報学習方法を提供する。
【解決手段】行動情報学習装置３００は、ロボット１００がワーク置き場から取得したワーク７を、作業者Ｐに対してワーク７を受け渡す領域である受渡領域８内で受け渡す場合に、ロボット１００の状態ｓを取得する状態情報取得部３１０と、状態ｓの調整情報である行動ａを出力する行動情報出力部３２０と、ワーク７の受け渡しにかかる受渡時間Ｔについての情報である判定情報を取得し、取得した判定情報に基づいて強化学習における報酬の値を算出する報酬算出部３３１と、報酬算出部３３１により算出された報酬の値と、状態ｓと、行動ａとに基づいて強化学習を行うことにより価値関数Ｑを更新する価値関数更新部３３２と、を備える。
【選択図】図２



【特許請求の範囲】
【請求項１】
ロボットがワーク置き場から取得したワークを、作業者に対してワークを受け渡す領域であるワーク受渡領域内で受け渡す場合に、前記ロボットの状態情報を取得する状態情報取得手段と、
前記状態情報の調整情報である行動情報を出力する行動情報出力手段と、
ワークの受け渡しにかかる受渡時間についての情報である判定情報を取得し、取得した前記判定情報に基づいて強化学習における報酬の値を算出する報酬算出手段と、
前記報酬算出手段により算出された前記報酬の値と、前記状態情報と、前記行動情報とに基づいて前記強化学習を行うことにより価値関数を更新する価値関数更新手段と、
を備える行動情報学習装置。
【請求項２】
請求項１に記載の行動情報学習装置において、
前記状態情報は、前記ロボットの姿勢及び前記ワーク受渡領域内の受渡位置に関する情報を含み、
前記調整情報は、前記状態情報についての調整を行うための情報を含むこと、
を特徴とする行動情報学習装置。
【請求項３】
請求項２に記載の行動情報学習装置において、
前記状態情報は、更にワークを取得した位置から前記ワーク受渡領域内への前記ロボットの移動経路を含むこと、
を特徴とする行動情報学習装置。
【請求項４】
請求項１から請求項３までのいずれかに記載の行動情報学習装置において、
前記報酬算出手段は、
前記受渡時間が前回の受渡時間よりも短い場合に、前記報酬の値を正の値とし、
前記受渡時間が前回の受渡時間よりも長い場合に、前記報酬の値を負の値とすること、
を特徴とする行動情報学習装置。
【請求項５】
請求項１から請求項４までのいずれかに記載の行動情報学習装置において、
前記受渡時間は、ワークを取得してから前記ワーク受渡領域内の位置に移動するまでの移動時間と、ワークを前記ワーク受渡領域内の位置に移動後、前記ワークを作業者が受け取るまでの開放時間とからなり、
前記受渡時間が同じ場合には、前記開放時間が短い場合に、前記移動時間が短い場合より前記報酬の値をより大きな値にすること、
を特徴とする行動情報学習装置。
【請求項６】
請求項１から請求項５までのいずれかに記載の行動情報学習装置において、
他の行動情報学習装置との間で前記価値関数を共有し、
前記価値関数更新手段が、前記共有した価値関数を更新すること、
を特徴とする行動情報学習装置。
【請求項７】
請求項１から請求項６までのいずれかに記載の行動情報学習装置と、
前記行動情報学習装置に対して通信ネットワークを介して接続され、前記ロボットを制御するロボット制御装置と、
を備えたロボット制御システムであって、
前記行動情報学習装置は、
前記価値関数更新手段により更新された前記価値関数に基づいて、前記ロボットによる前記受渡時間を最短にするための行動情報である最適化行動情報を生成する行動情報生

成手段と、
　前記行動情報生成手段により生成された前記最適化行動情報を、前記ロボット制御装置に対して出力する行動情報出力手段と、
　を備えるロボット制御システム。
【請求項８】
　状態情報取得手段が、ロボットがワーク置き場から取得したワークを、作業者に対してワークを受け渡す領域であるワーク受渡領域内で受け渡す場合に、前記ロボットの状態情報を取得するステップと、
　行動情報出力手段が、前記状態情報の調整情報である行動情報を出力するステップと、
　報酬算出手段が、ワークの受け渡しにかかる受渡時間についての情報である判定情報を取得し、取得した前記判定情報に基づいて強化学習における報酬の値を算出するステップと、
　価値関数更新手段が、算出された前記報酬の値と、前記状態情報と、前記行動情報とに基づいて前記強化学習を行うことにより価値関数を更新するステップと、
　を含む行動情報学習方法。
【発明の詳細な説明】
【技術分野】
【０００１】
　本発明は、行動情報学習装置、ロボット制御システム及び行動情報学習方法に関する。
【背景技術】
【０００２】
　従来、ロボットが人と共存する空間で作業する場合がある。例えば、ロボットは、予めプログラミングされた位置までワークを運び、作業者にワークを受け渡すと、次のワークを取りに行くという一連の動作を繰り返す場合である。
　このように、ロボットと作業者との間で協調作業を行う場合、ロボットは、プログラミングによって定められた位置やタイミング等で作業を行っていた。しかし、このような場合に、受け渡し位置やタイミングによっては、作業者にとって受け取り難い場合や、次の作業を始め難い場合がある。
　ここで、ロボットにさせる作業を最適化するための装置が開示されている（例えば、特許文献１参照）。
【先行技術文献】
【特許文献】
【０００３】
【特許文献１】特開２００９－１２５９２０号公報
【発明の概要】
【発明が解決しようとする課題】
【０００４】
　ロボットに最適な処理をさせるためには、再度プログラミングを行って受け渡し位置等を教示する必要があるが、試行錯誤して最適な位置等を探し出すのには限度があった。
【０００５】
　本発明は、係る課題に鑑みなされたものであり、その目的は、作業者がロボットとの協調作業を行いやすくするための行動情報学習装置、ロボット制御システム及び行動情報学習方法を提供することである。
【課題を解決するための手段】
【０００６】
　（１）本発明の行動情報学習装置（例えば、行動情報学習装置３００）は、ロボット（例えば、ロボット１００）がワーク置き場から取得したワーク（例えば、ワーク７）を、作業者（例えば、作業者Ｐ）に対してワークを受け渡す領域であるワーク受渡領域（例えば、受渡領域８）内で受け渡す場合に、前記ロボットの状態情報（例えば、状態ｓ）を取得する状態情報取得手段（例えば、状態情報取得部３１０）と、前記状態情報の調整情報

である行動情報（例えば、行動ａ）を出力する行動情報出力手段（例えば、行動情報出力部３２０）と、ワークの受け渡しにかかる受渡時間（例えば、受渡時間Ｔ）についての情報である判定情報を取得し、取得した前記判定情報に基づいて強化学習における報酬の値を算出する報酬算出手段（例えば、報酬算出部３３１）と、前記報酬算出手段により算出された前記報酬の値と、前記状態情報と、前記行動情報とに基づいて前記強化学習を行うことにより価値関数（例えば、価値関数Ｑ）を更新する価値関数更新手段（例えば、価値関数更新部３３２）と、を備える。

【０００７】

（２）　（１）に記載の行動情報学習装置において、前記状態情報は、前記ロボットの姿勢及び前記ワーク受渡領域内の受渡位置に関する情報を含み、前記調整情報は、前記状態情報についての調整を行うための情報を含んでもよい。

【０００８】

（３）　（２）に記載の行動情報学習装置において、前記状態情報は、更にワークを取得した位置から前記ワーク受渡領域内への前記ロボットの移動経路を含んでもよい。

【０００９】

（４）　（１）から（３）までのいずれかに記載の行動情報学習装置において、前記報酬算出手段は、前記受渡時間が前回の受渡時間よりも短い場合に、前記報酬の値を正の値とし、前記受渡時間が前回の受渡時間よりも長い場合に、前記報酬の値を負の値としてもよい。

【００１０】

（５）　（１）から（４）までのいずれかに記載の行動情報学習装置において、前記受渡時間は、ワークを取得してから前記ワーク受渡領域内の位置に移動するまでの移動時間（例えば、移動時間Ｔ１）と、ワークを前記ワーク受渡領域内の位置に移動後、前記ワークを作業者が受け取るまでの開放時間（例えば、開放時間Ｔ２）とからなり、前記受渡時間が同じ場合には、前記開放時間が短い場合に、前記移動時間が短い場合より前記報酬の値をより大きな値にしてもよい。

【００１１】

（６）　（１）から（５）までのいずれかに記載の行動情報学習装置において、他の行動情報学習装置との間で前記価値関数を共有し、前記価値関数更新手段が、前記共有した価値関数を更新してもよい。

【００１２】

（７）　本発明によるロボット制御システム（例えば、ロボット制御システム１０００）は、（１）から（５）までのいずれかに記載の行動情報学習装置（例えば、行動情報学習装置３００）と、前記行動情報学習装置に対して通信ネットワーク（例えば、ネットワーク４００）を介して接続され、前記ロボット（例えば、ロボット１００）を制御するロボット制御装置（例えば、ロボット制御装置２００）と、を備え、前記行動情報学習装置が、前記価値関数更新手段（例えば、価値関数更新部３３２）により更新された前記価値関数に基づいて、前記ロボットによる前記受渡時間を最短にするための行動情報である最適化行動情報を生成する行動情報生成手段（例えば、最適化行動情報出力部３５０）と、前記行動情報生成手段により生成された前記最適化行動情報を、前記ロボット制御装置に対して出力する行動情報出力手段（例えば、最適化行動情報出力部３５０）と、を備える。

【００１３】

（８）　本発明による行動情報学習方法は、状態情報取得手段が、ロボットがワーク置き場から取得したワークを、作業者に対してワークを受け渡す領域であるワーク受渡領域内で受け渡す場合に、前記ロボットの状態情報を取得するステップと、行動情報出力手段が、前記状態情報の調整情報である行動情報を出力するステップと、報酬算出手段が、ワークの受け渡しにかかる受渡時間についての情報である判定情報を取得し、取得した前記判定情報に基づいて強化学習における報酬の値を算出するステップと、価値関数更新手段が、算出された前記報酬の値と、前記状態情報と、前記行動情報とに基づいて前記強化学

習を行うことにより価値関数を更新するステップと、を含む。
【発明の効果】
【００１４】
本発明によれば、作業者がロボットとの協調作業を行いやすくするための行動情報学習装置、ロボット制御システム及び行動情報学習方法を提供できる。
【図面の簡単な説明】
【００１５】
【図１】本発明の実施形態全体の基本的構成を示すブロック図である。
【図２】本発明の実施形態に含まれる各装置が備える機能ブロックを表すブロック図である。
【図３】本発明の実施形態におけるロボットの動作を説明するための図である。
【図４】本発明の実施形態における強化学習時の基本的動作を示すフローチャートである。
【図５】本発明の実施形態における最適化行動情報の選択時の基本的動作を示すフローチャートである。
【図６】本発明の変形形態における行動情報学習装置間の連携を示すブロック図である。
【発明を実施するための形態】
【００１６】
（実施形態）
まず、本発明の実施形態の概略を説明する。本実施形態において、図１に示すように、ロボット１００と、作業者Ｐとは、作業空間内にて共同で作業をする。そして、ロボット１００は、ワーク７（図３参照）を作業者Ｐに運搬し、作業者Ｐは、ロボット１００からワーク７を受け取って作業をする。そして、ロボット制御システム１０００では、ロボット１００がワーク７を運搬してから、作業者Ｐがワーク７を受け取るまでの時間が最短になるように、ロボット１００の行動情報を学習する。
【００１７】
次に、本実施形態に係るロボット制御システム１０００の構成について説明する。ロボット制御システム１０００は、ロボット１００、ロボット制御装置２００、行動情報学習装置３００及びネットワーク４００を備えている。
ここで、ロボット制御装置２００とロボット１００とは、１対１の組とされて、通信可能に接続されている。なお、ロボット制御装置２００とロボット１００とは、接続インタフェースを介して直接接続されても、また、ＬＡＮ（Ｌｏｃａｌ Ａｒｅａ Ｎｅｔｗｏｒｋ）等のネットワークを介して接続されてもよい。
【００１８】
また、ロボット制御装置２００と、行動情報学習装置３００とは、それぞれ接続インタフェースを介して直接に接続、又は、それぞれネットワーク４００を介して接続されており、相互に通信を行うことが可能である。なお、ネットワーク４００は、例えば、工場内に構築されたＬＡＮや、インターネット、公衆電話網、或いは、これらの組み合わせである。ネットワーク４００における具体的な通信方式や、有線接続及び無線接続のいずれであるか等については、特に限定されない。
【００１９】
次に、ロボット制御システム１０００に含まれるこれら装置の機能について、図２を参照して説明する。ここで、図２は、各装置に含まれる機能ブロックを表すブロック図である。なお、各装置間に存在するネットワーク４００については、その図示を省略する。
【００２０】
ロボット１００は、ロボット制御装置２００に設定されたロボット制御プログラム及びロボット制御装置２００に設定されたパラメータの設定値に基づいて生成される動作指令にしたがって、例えば、部品等のワーク７を運搬する。ロボット１００は、モータ部１１０と、開放ボタン１２０とを備える。
モータ部１１０は、ロボット１００のハンド部１３（後述する）等の駆動軸を駆動させ

るサーボモータである。
開放ボタン１２０は、ハンド部１３に把持したワーク７を取り外す処理を行うためのボタンである。開放ボタン１２０は、作業者Ｐにより操作される。開放ボタン１２０を操作したことによる押下情報は、ロボット制御装置２００に送られる。
【００２１】
ここで、ロボット１００による動作について、図３に基づき説明する。
図３は、本発明の実施形態におけるロボット１００の動作を説明するための図である。
ロボット１００は、例えば、６軸多関節型のロボットである。ロボット１００の各関節部の駆動軸及びハンド部１３の駆動軸は、モータ部１１０によって駆動するが、ロボット制御装置２００によって制御される。
ロボット１００は、例えば、ワーク置き場に載置されたワーク７を取得し、作業台上の受渡領域８の所定位置にワーク７を運搬する。このロボット１００がワーク７を取得してから受渡領域８の所定位置までワーク７を運搬するまでの時間を、移動時間Ｔ１とする。
【００２２】
作業者Ｐによるロボット１００への操作、例えば、作業者Ｐがロボット１００のハンド部１３の近傍を掴んで動かす動作をすることによって、ロボット１００は、位置及び姿勢を変える。また、作業者Ｐによる開放ボタン１２０の押下操作によって、ロボット制御装置２００のモータ制御部２１０は、ワーク７をハンド部１３から取り外す制御を行い、作業者Ｐは、ロボット１００からワーク７を受け取る。このロボット１００がワーク７を受渡領域８の所定位置まで運搬してから作業者Ｐがワーク７を受け取るまでの時間を、開放時間Ｔ２とする。そして、移動時間Ｔ１と、開放時間Ｔ２とをあわせた時間を、受渡時間Ｔとする。
【００２３】
以上、ロボット１００の機能ブロックについて説明したが、上述した機能ブロックは、本実施形態の動作に特に関連する部分である。ロボット１００は、上述した機能ブロック以外にも、例えば、動作指令を増幅するモータ駆動アンプや、ユーザの操作を受け付けるための操作盤等、一般的な機能ブロックを備えている。しかしながら、これらの一般的な機能ブロックについては、当業者によく知られているので、詳細な説明及び図示を省略する。
【００２４】
図２に戻り、ロボット制御装置２００は、ロボット１００を制御することにより、ロボット１００に所定の動作を行わせる装置である。また、ロボット制御装置２００は、行動情報学習装置３００に対して状態情報（「ステータス」ともいう。）を送信する。更に、ロボット制御装置２００は、行動情報学習装置３００から行動情報（「アクション」ともいう。）を受信する。これら各情報の詳細については、行動情報学習装置３００の機能ブロックの説明と併せて説明をする。
【００２５】
ロボット制御装置２００は、モータ制御部２１０と、パラメータ設定部２２０と、プログラム修正部２３０と、受渡時間計測部２４０とを備える。
【００２６】
モータ制御部２１０は、ロボット制御プログラム及びパラメータ（例えば、ハンド部１３を含むロボット１００の受渡領域８内での姿勢、受渡領域８内における位置、ワーク７を取得してから受渡領域８内の位置に達するまでのロボット１００の移動経路に関する値）の設定値に基づいて動作指令を生成し、生成した動作指令をロボット１００に送出する。そして、モータ制御部２１０は、ロボット１００に動作指令を送出することにより、ロボット１００のモータ部１１０等の駆動を制御する。この処理により、ロボット１００によるワーク７の運搬動作が実現される。ここで、ロボット制御プログラムには、運搬のための諸条件（例えば、障害物を避けるためのマップ情報、移動速度等）が定義されている。
【００２７】

パラメータ設定部２２０は、ロボット１００の当該ロボット制御プログラムによるワーク７の運搬時における移動処理に関するパラメータを設定する部分である。ロボット１００のパラメータとは、例えば、ハンド部１３を含むロボット１００の受渡領域８内での姿勢や、受渡領域８内の位置や、ロボット１００の移動経路に関するパラメータである。ハンド部１３を含むロボット１００の姿勢を示すパラメータは、例えば、ハンド部１３の角度に関するデータである。また、受渡領域８内の位置を示すパラメータは、例えば、ハンド部１３の位置をＸＹＺ座標で表したデータである。ロボット１００の移動経路に関するパラメータは、例えば、ロボット１００がワーク７を取得したワーク置き場の位置から受渡領域８までのロボット１００の教示点のデータである。ここで、ロボット１００の教示点とは、ロボット１００のハンド部１３の先端部の位置をいう。

かかるパラメータの設定値は、行動情報学習装置３００から出力される行動情報や、最適化行動情報に基づいて調整される。

【００２８】

プログラム修正部２３０は、ロボット制御プログラムを直接修正する。具体的には、プログラム修正部２３０は、当該ロボット制御プログラムで記述されたハンド部１３を含むロボット１００の受渡領域８内での姿勢や、受渡領域８内の位置等を、行動情報学習装置３００から出力される行動情報や、最適化行動情報に基づいて、プログラムコードを直接修正する。

【００２９】

受渡時間計測部２４０は、時間を計測する制御部である。受渡時間計測部２４０は、ロボット１００がワーク置き場に載置されたワーク７を取得し、受渡領域８の所定位置にワーク７を運搬するまでの時間である移動時間Ｔ１を計測する。また、受渡時間計測部２４０は、受渡領域８の所定位置にワーク７を運搬してから作業者Ｐがワーク７を受け取るまでの時間である開放時間Ｔ２を計測する。

【００３０】

行動情報学習装置３００は、強化学習を行う装置である。行動情報学習装置３００に含まれる各機能ブロックの説明に先立って、まずは、強化学習の基本的な仕組みについて説明する。エージェント（本実施形態における行動情報学習装置３００に相当）は、環境の状態を観測し、ある行動を選択し、当該行動に基づいて環境が変化する。環境の変化に伴って、何らかの報酬が与えられ、エージェントはより良い行動の選択（意思決定）を学習する。

教師あり学習が、完全な正解を示すのに対して、強化学習における報酬は、環境の一部の変化に基づく断片的な値であることが多い。

このため、エージェントは、将来にわたっての報酬の合計を最大にするように行動を選択するように学習する。

【００３１】

このように、強化学習では、行動を学習することにより、環境に行動が与える相互作用を踏まえて適切な行動を学習、すなわち将来的に得られる報酬を最大にするための学習する方法を学ぶ。これは、本実施形態において、例えば、受渡時間Ｔを短縮し、更に、開放時間Ｔ２を短縮するための行動情報を選択するという、未来に影響をおよぼすような行動を獲得できることを表している。

【００３２】

ここで、強化学習としては、任意の学習方法を用いることができるが、以下の説明では、或る環境の状態ｓの下で、行動ａを選択する価値Ｑ（ｓ，ａ）を学習する方法であるＱ学習（Ｑ－ｌｅａｒｎｉｎｇ）を用いる場合を例にとって説明をする。

Ｑ学習では、或る状態ｓのとき、取り得る行動ａのなかから、価値Ｑ（ｓ，ａ）の最も高い行動ａを最適な行動として選択することを目的とする。

【００３３】

しかしながら、Ｑ学習を最初に開始する時点では、状態ｓと行動ａとの組み合わせについて、価値Ｑ（ｓ，ａ）の正しい値は全く分かっていない。そこで、エージェントは、或

る状態ｓの下で様々な行動ａを選択し、その時の行動ａに対して、与えられる報酬に基づいて、より良い行動の選択をすることにより、正しい価値Ｑ（ｓ，ａ）を学習していく。

【００３４】

また、将来にわたって得られる報酬の合計を最大化したいので、最終的に価値Ｑ（ｓ，ａ）＝Ｅ［Σ（γ$^{t}$）ｒ$_{t}$］となるようにすることを目指す。ここでＥ［］は期待値を表し、ｔは時刻、γは後述する割引率と呼ばれるパラメータ、ｒ$_{t}$は時刻ｔにおける報酬、Σは時刻ｔによる合計である。この式における期待値は、最適な行動にしたがって状態変化した場合の期待値である。しかしＱ学習の過程において、最適な行動が何であるのかは不明であるので、様々な行動を行うことにより、探索しながら強化学習をする。このような価値Ｑ（ｓ，ａ）の更新式は、例えば、次の式（１）により表すことができる。

【００３５】

【数１】

$$Q(s_{t+1},a_{t+1}) \leftarrow Q(s_t,a_t)+\alpha(r_{t+1}+\gamma \max_a Q(s_{t+1},a)-Q(s_t,a_t)) \quad \cdots(1)$$

【００３６】

上記の式（１）において、ｓ$_{t}$は、時刻ｔにおける環境の状態を表し、ａ$_{t}$は、時刻ｔにおける行動を表す。行動ａ$_{t}$により、状態はｓ$_{t+1}$に変化する。ｒ$_{t+1}$は、その状態の変化により得られる報酬を表している。また、ｍａｘの付いた項は、状態ｓ$_{t+1}$の下で、その時に分かっている最もＱ値の高い行動ａを選択した場合のＱ値にγを乗じたものになる。ここで、γは、０＜γ≦１のパラメータで、割引率と呼ばれる。また、αは、学習係数で、０＜α≦１の範囲とする。

【００３７】

上述した式（１）は、試行ａ$_{t}$の結果、返ってきた報酬ｒ$_{t+1}$を元に、状態ｓ$_{t}$における行動ａ$_{t}$の価値Ｑ（ｓ$_{t}$，ａ$_{t}$）を更新する方法を表している。　この更新式は、状態ｓ$_{t}$における行動ａ$_{t}$の価値Ｑ（ｓ$_{t}$，ａ$_{t}$）よりも、行動ａ$_{t}$による次の状態ｓ$_{t+1}$における最良の行動の価値ｍａｘ$_{a}$　Ｑ（ｓ$_{t+1}$，ａ）の方が大きければ、価値Ｑ（ｓ$_{t}$，ａ$_{t}$）を大きくし、逆に小さければ、価値Ｑ（ｓ$_{t}$，ａ$_{t}$）を小さくすることを示している。つまり、或る状態における或る行動の価値を、それによる次の状態における最良の行動の価値に近づける。ただし、その差は、割引率γと報酬ｒ$_{t+1}$のあり方により変わってくるが、基本的には、ある状態における最良の行動の価値が、それに至る一つ前の状態における行動の価値に伝播していく仕組みになっている。

【００３８】

ここで、Ｑ学習では、すべての状態行動ペア（ｓ，ａ）についての価値Ｑ（ｓ，ａ）のテーブルを作成して、学習を行う方法がある。しかし、すべての状態行動ペアの価値Ｑ（ｓ，ａ）の値を求めるには状態数が多すぎて、Ｑ学習が収束するのに多くの時間を要してしまう場合がある。

【００３９】

そこで、公知のＤＱＮ（Ｄｅｅｐ　Ｑ－Ｎｅｔｗｏｒｋ）と呼ばれる技術を利用するようにしてもよい。具体的には、価値関数Ｑを適当なニューラルネットワークを用いて構成し、ニューラルネットワークのパラメータを調整することにより、価値関数Ｑを適当なニューラルネットワークで近似することにより価値Ｑ（ｓ，ａ）の値を算出するようにしてもよい。ＤＱＮを利用することにより、Ｑ学習が収束するのに要する時間を短くすることが可能となる。なお、ＤＱＮについては、例えば、以下の非特許文献に詳細な記載がある。

【００４０】

＜非特許文献＞

「Ｈｕｍａｎ－ｌｅｖｅｌ　ｃｏｎｔｒｏｌ　ｔｈｒｏｕｇｈ　ｄｅｅｐ　ｒｅｉｎｆｏｒｃｅｍｅｎｔ　ｌｅａｒｎｉｎｇ」、Ｖｏｌｏｄｙｍｙｒ　Ｍｎｉｈ１著［ｏｎｌｉｎｅ］、［平成２９年３月１７日検索］、インターネット〈ＵＲＬ：ｈｔｔｐ：／／ｆｉ

ｌｅｓ．ｄａｖｉｄｑｉｕ．ｃｏｍ／ｒｅｓｅａｒｃｈ／ｎａｔｕｒｅ１４２３６．ｐｄｆ〉

【００４１】

行動情報学習装置３００は、上記において説明をしたＱ学習を行う。具体的には、行動情報学習装置３００は、ロボット１００において設定されたロボット制御プログラムの内容及び当該ロボット制御プログラム実行時におけるパラメータの組み合わせを状態ｓとし、当該状態ｓに係る当該ロボット制御プログラムの修正及びパラメータの調整を行動ａとして、選択する価値関数Ｑを学習する。

【００４２】

行動情報学習装置３００は、ロボット１００において設定されたロボット制御プログラム及びパラメータ等の状態ｓを観測して、行動ａを決定する。行動情報学習装置３００は、行動ａをするたびに報酬が返ってくる。行動情報学習装置３００は、将来にわたっての報酬の合計が最大になるように、最適な行動ａを試行錯誤的に探索する。そうすることで、行動情報学習装置３００は、ロボット１００において設定されたロボット制御プログラムの内容及び当該ロボット制御プログラム実行時におけるパラメータの組み合わせ等である状態ｓに対して、最適な行動ａを選択することが可能となる。

【００４３】

すなわち、行動情報学習装置３００により学習された価値関数Ｑに基づいて、或る状態ｓに係るロボット制御プログラムの内容及び当該ロボット制御プログラム実行時におけるパラメータの組み合わせに対して、適用される行動ａのうち、価値関数Ｑの値が最大となるような行動ａを選択することで、ワーク７の受け渡しに係る時間である受渡時間Ｔ及び開放時間Ｔ２が最短になるような行動ａを選択することが可能となる。

【００４４】

以上の強化学習を行うために、行動情報学習装置３００は、状態情報取得部３１０、行動情報出力部３２０、学習部３３０及び価値関数記憶部３４０を備える。

【００４５】

状態情報取得部３１０は、ロボット制御プログラムの内容及び当該ロボット制御プログラム実行時におけるパラメータの組み合わせである状態情報（状態ｓ）を、ロボット制御装置２００（及び／又はロボット１００）から取得する部分である。この状態ｓは、Ｑ学習における、環境の状態ｓに相当する。

【００４６】

具体的には、本実施形態における状態ｓには、ロボット１００を制御するためのロボット制御プログラムの内容及び当該ロボット制御プログラム実行時におけるパラメータの組み合わせが含まれ、パラメータには、当該ロボット制御プログラム等によるハンド部１３を含むロボット１００の受渡領域８内での姿勢、受渡領域８内における位置、ワーク７を取得してから受渡領域８内の位置に達するまでの移動経路に関する情報が含まれる。状態情報取得部３１０は、取得した状態ｓを学習部３３０に対して出力する。

【００４７】

また、状態情報取得部３１０は、Ｑ学習を行うための報酬を算出するための判定情報も取得する。具体的には、状態ｓに係るロボット制御プログラム及び当該ロボット制御プログラム実行時におけるパラメータの組み合わせを実行後の受渡時間Ｔを、Ｑ学習を行うための報酬を算出するための判定情報とする。受渡時間Ｔは、上述したように、ロボット１００がワーク７を取得してから受渡領域８内の位置まで移動する時間である移動時間Ｔ１と、受渡領域８内の位置に移動してから作業者Ｐにワーク７が受け渡されるまでの開放時間Ｔ２とからなる。

【００４８】

行動情報出力部３２０は、学習部３３０が生成した行動情報（行動ａ）を、ロボット制御装置２００に対して送信する部分である。ロボット制御装置２００は、上述したように、この行動ａに基づいて、現在の状態ｓ、すなわち現在設定されているロボット制御プログラム及びパラメータを修正することで、次の状態ｓ′（すなわち修正されたロボット制

御プログラム、修正されたパラメータ及び当該修正されたロボット制御プログラムと修正されたパラメータによる移動処理を実行した場合の状態）に遷移する。

【００４９】

学習部３３０は、或る環境の状態ｓの下で、ある行動ａを選択する場合の価値Ｑ（ｓ，ａ）を学習する部分である。具体的には、学習部３３０は、報酬算出部３３１、価値関数更新部３３２及び行動情報生成部３３３を備える。

【００５０】

報酬算出部３３１は、或る状態ｓの下で、行動ａを選択した場合の報酬を、判定情報に基づいて算出する部分である。ここで、本実施形態では、行動ａにより修正された状態ｓ´に係る修正後のロボット制御プログラム及び当該修正後のロボット制御プログラム実行時における修正後のパラメータに基づいて動作したロボット１００の受渡時間Ｔが、行動ａにより修正される前の状態ｓに係る修正前のロボット制御プログラム及び当該修正前のロボット制御プログラム実行時における修正前のパラメータに基づいて動作したロボット１００の受渡時間Ｔよりも長くなった場合に、報酬の値を負の値とする。

【００５１】

一方で、行動ａにより修正された状態ｓ´に係る修正後のロボット制御プログラム及び当該修正後のロボット制御プログラム実行時における修正後のパラメータに基づいて動作したロボット１００の受渡時間Ｔが、行動ａにより修正される前の状態ｓに係る修正前のロボット制御プログラム及び当該修正前のロボット制御プログラム実行時における修正前のパラメータに基づいて動作したロボット１００の受渡時間Ｔよりも短くなった場合に、報酬の値を正の値とする。

【００５２】

また、報酬の値については重みづけを与えるようにすることができる。例えば、移動時間Ｔ１と、開放時間Ｔ２とでは、開放時間Ｔ２が短くなった場合の方が、移動時間Ｔ１が短くなった場合に比べて、報酬の正の値を大きくすることが好ましい。つまり、開放時間Ｔ２が短くなった度合いに応じて、正の値が大きくなるようにするとよい。

なお、上記の報酬の値の算出方法は、一例であって、これに限定されない。例えば、状態ｓ´と状態ｓにおける受渡時間Ｔの偏差、移動時間Ｔ１の偏差、及び開放時間Ｔ２の偏差と、報酬の値を対応付ける報酬対応テーブル（仮称）を予め任意に作成しておき、報酬対応テーブルに基づいて、報酬の値を算出するようにしてもよい。また、受渡時間Ｔの偏差、移動時間Ｔ１の偏差、及び開放時間Ｔ２の偏差を入力とする報酬関数（仮称）を予め任意に作成しておき、報酬関数に基づいて、報酬の値を算出するようにしてもよい。

【００５３】

価値関数更新部３３２は、状態ｓと、行動ａと、行動ａを状態ｓに適用した場合の状態ｓ´と、上記のようにして算出された報酬の値とに基づいてＱ学習を行うことにより、価値関数記憶部３４０が記憶する価値関数Ｑを更新する。

【００５４】

価値関数Ｑの更新は、オンライン学習で行ってもよく、バッチ学習で行ってもよく、ミニバッチ学習で行ってもよい。

オンライン学習とは、或る行動ａを現在の状態ｓに適用することにより、状態ｓが新たな状態ｓ´に遷移する都度、即座に価値関数Ｑの更新を行うという学習方法である。また、バッチ学習とは、或る行動ａを現在の状態ｓに適用することにより、状態ｓが新たな状態ｓ´に遷移することを繰り返すことにより、学習用のデータを収集し、収集したすべての学習用データを用いて、価値関数Ｑの更新を行うという学習方法である。更に、ミニバッチ学習とは、オンライン学習と、バッチ学習の中間的な、ある程度学習用データが溜まるたびに価値関数Ｑの更新を行うという学習方法である。

【００５５】

行動情報生成部３３３は、Ｑ学習の過程において、ロボット１００に様々な動作（Ｑ学習における行動ａに相当する。）を行わせるために、行動ａを生成して、生成した行動ａを行動情報出力部３２０に対して出力する。

【００５６】
具体的には、行動情報生成部３３３は、現在の状態ｓに対して、Ｑ学習の過程における行動ａを選択する。本実施形態における行動ａには、現在の状態ｓに係るロボット制御プログラムで記述された内容に関する修正情報、及び現在の状態ｓに係るパラメータ（例えば、ハンド部１３を含むロボット１００の姿勢、受渡領域８内における位置、ワーク７を取得してから受渡領域８内の位置に達するまでのロボット１００の移動経路に関する値）の設定値が含まれる。
【００５７】
行動情報生成部３３３は、例えば、状態ｓに含まれるロボット制御プログラム及びパラメータに対して行動ａに含まれるパラメータ（例えば、ハンド部１３を含むロボット１００の姿勢、受渡領域８内における位置、ワーク７を取得してから受渡領域８内の位置に達するまでのロボット１００の移動経路に関する値）の設定値を適用して、状態ｓ′に遷移して、プラスの報酬（正の値の報酬）が返った場合、次の行動ａ′としては、例えば、受渡領域８内の位置を、ワーク７を取り外した位置である開放位置側に少し移動させたり、ロボット１００の姿勢を、ワーク７を取り外した姿勢である開放姿勢の方向に少し変化させたりして、受渡時間Ｔがより短くなるような行動ａ′を選択する方策を取るようにしてもよい。
【００５８】
また、逆に、マイナスの報酬（負の値の報酬）が返った場合、行動情報生成部３３３は、例えば、状態ｓ′よりも状態ｓ寄りになるように、行動ａ′を選択するようにしてもよい。又は、状態ｓ′寄りになるような行動ａ′を選択することで、マイナスの報酬になると思われる行動を集めるようにしてもよい。
更に、行動情報生成部３３３は、現在の推定される行動ａの価値の中で、最も価値Ｑ（ｓ，ａ）の高い行動ａ′を選択するグリーディ法や、ある小さな確率εでランダムに行動ａ′選択し、それ以外では最も価値Ｑ（ｓ，ａ）の高い行動ａ′を選択するεグリーディ法といった公知の方法により、行動ａ′を選択する方策を取るようにしてもよい。
【００５９】
価値関数記憶部３４０は、価値関数Ｑを記憶する記憶装置である。価値関数記憶部３４０に記憶された価値関数Ｑは、価値関数更新部３３２により更新される。
【００６０】
また、行動情報学習装置３００は、価値関数更新部３３２がＱ学習を行うことにより更新した価値関数Ｑに基づいて、価値Ｑ（ｓ，ａ）が最大となる動作をロボット１００に行わせるための行動ａ（以下、「最適化行動情報」と呼ぶ。）を生成する。
【００６１】
行動情報学習装置３００は、最適化行動情報出力部３５０を備えている。
最適化行動情報出力部３５０は、価値関数記憶部３４０が記憶している価値関数Ｑを取得する。この価値関数Ｑは、上述したように、価値関数更新部３３２がＱ学習を行うことにより更新したものである。そして、最適化行動情報出力部３５０は、価値関数Ｑに基づいて、最適化行動情報を生成し、生成した最適化行動情報をロボット制御装置２００に対して出力する。この最適化行動情報には、行動情報出力部３２０がＱ学習の過程において出力する行動情報と同様に、修正後のロボット制御プログラム及び当該修正後のロボット制御プログラム実行時における修正後のパラメータが含まれる。
【００６２】
ロボット制御装置２００が、この最適化行動情報に基づいて現在設定されているロボット制御プログラム及びパラメータを修正して、動作指令を生成することにより、ロボット１００は、受渡時間Ｔ及び開放時間Ｔ２が最短になるように動作することができる。
【００６３】
以上、ロボット制御装置２００や行動情報学習装置３００に含まれる機能ブロックについて説明した。
これらの機能ブロックを実現するために、ロボット制御装置２００及び行動情報学習装

置３００は、ＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）等の演算処理装置を備える。また、ロボット制御装置２００及び行動情報学習装置３００は、アプリケーションソフトウェアやＯＳ（Ｏｐｅｒａｔｉｎｇ　Ｓｙｓｔｅｍ）等の各種の制御用プログラムを格納したＨＤＤ（Ｈａｒｄ　Ｄｉｓｋ　Ｄｒｉｖｅ）等の補助記憶装置や、演算処理装置がプログラムを実行する上で一時的に必要とされるデータを格納するためのＲＡＭ（Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ）といった主記憶装置も備える。

【００６４】

そして、ロボット制御装置２００及び行動情報学習装置３００は、演算処理装置が補助記憶装置からアプリケーションソフトウェアやＯＳを読み込み、読み込んだアプリケーションソフトウェアやＯＳを主記憶装置に展開させながら、これらのアプリケーションソフトウェアやＯＳに基づいた演算処理を行う。また、ロボット制御装置２００及び行動情報学習装置３００は、この演算結果に基づいて、各装置が備える各種のハードウェアを制御する。これにより、本実施形態の機能ブロックは実現される。つまり、本実施形態は、ハードウェアとソフトウェアが協働することにより実現することができる。

【００６５】

具体例として、ロボット制御装置２００は、一般的なロボット１００の制御装置に本実施形態を実現するためのアプリケーションソフトウェアを組み込むことにより実現できる。また、行動情報学習装置３００は、一般的なパーソナルコンピュータに、本実施形態を実現するためのアプリケーションソフトウェアを組み込むことより実現できる。

【００６６】

ただし、行動情報学習装置３００については、機械学習に伴う演算量が多いため、例えば、パーソナルコンピュータにＧＰＵ（Ｇｒａｐｈｉｃｓ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔｓ）を搭載し、ＧＰＧＰＵ（Ｇｅｎｅｒａｌ－Ｐｕｒｐｏｓｅ　ｃｏｍｐｕｔｉｎｇ　ｏｎ　Ｇｒａｐｈｉｃｓ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔｓ）と呼ばれる技術により、ＧＰＵを機械学習に伴う演算処理に利用するようにすると、高速処理できるようになるのでよい。更には、より高速な処理を行うために、行動情報学習装置３００は、このようなＧＰＵを搭載したコンピュータを複数台用いてコンピュータ・クラスターを構築し、このコンピュータ・クラスターに含まれる複数のコンピュータにて並列処理を行うようにしてもよい。

【００６７】

次に、図４のフローチャートを参照して本実施形態における行動情報学習処理として、Ｑ学習を行う行動情報学習装置３００の動作について説明をする。

【００６８】

まず、ステップＳ１１（以下、単に「Ｓ」という。）において、状態情報取得部３１０がロボット制御装置２００から状態情報を取得する。取得した状態情報は、価値関数更新部３３２や行動情報生成部３３３に対して出力される。上述したように、この状態情報は、Ｑ学習における環境の状態ｓに相当する情報であり、Ｓ１１時点での、ロボット制御プログラムの内容とパラメータの設定値である、ハンド部１３を含むロボット１００の受渡領域８内での姿勢、受渡領域８内の位置、移動経路に関する情報が含まれる。なお、最初にＱ学習を開始する時点でのロボット制御プログラム及びパラメータの設定値は、予めユーザが生成するようにする。つまり、本実施形態では、ユーザが作成したロボット制御プログラム及びパラメータの初期設定値を、強化学習により最適なものに調整する。

【００６９】

Ｓ１２において、行動情報生成部３３３が新たな行動情報を生成し、生成した新たな行動情報（行動ａ）を、行動情報出力部３２０を介してロボット制御装置２００に対して出力する。行動情報を受信したロボット制御装置２００は、受信した行動情報に基づいて現在の状態ｓに係るロボット制御プログラム及びパラメータを修正した状態ｓ´により、ロボット１００を駆動させて、ワーク７の受け渡し処理を行う。上述したように、この行動情報は、Ｑ学習における行動ａに相当するものである。ここで、行動情報には、例えば、ロボット制御プログラムの修正値と、パラメータの設定値が含まれる点については、上述

した通りである。

【００７０】

Ｓ１３において、状態情報取得部３１０は、新たな状態ｓ´についての判定情報を取得する。ここで、新たな状態ｓ´には、状態ｓ´に係るロボット制御プログラム及びパラメータを含む。また、判定情報は、状態ｓ´に係る移動処理行うために要した移動時間Ｔ１及び開放時間Ｔ２からなる受渡時間Ｔを含む。取得した判定情報は、報酬算出部３３１に対して出力される。

【００７１】

報酬算出部３３１は、入力された判定情報に基づいて報酬を算出する。そのために、Ｓ１４において、報酬算出部３３１は、判定情報に含まれる受渡時間Ｔが短くなったか否かを判定する。かかる判定は、状態ｓ´の判定情報に含まれる、状態ｓ´に係る移動処理を行うために要した受渡時間Ｔと、状態ｓ´の前の状態である状態ｓの判定情報に含まれる、状態ｓに係る移動処理を行うために要した受渡時間Ｔとを比較することにより行うことができる。受渡時間Ｔが短くなったのであれば（Ｓ１４：Ｙｅｓ）、報酬算出部３３１は、処理をＳ１５に移す。他方、受渡時間Ｔが長くなったのであれば（Ｓ１４：ＮＯ）、報酬算出部３３１は、処理をＳ１８に移す。

【００７２】

Ｓ１５において、報酬算出部３３１は、判定情報に含まれる開放時間Ｔ２が、状態ｓ´の前の状態である状態ｓの判定情報に含まれる、状態ｓに係る移動処理を行うために要した開放時間Ｔ２より短くなったか否かを判定する。開放時間Ｔ２が短くなったのであれば（Ｓ１５：Ｙｅｓ）、報酬算出部３３１は、処理をＳ１６に移す。他方、開放時間Ｔ２が長くなったのであれば（Ｓ１５：ＮＯ）、報酬算出部３３１は、処理をＳ１７に移す。

【００７３】

Ｓ１６において、報酬算出部３３１は、報酬を第１の値とする。ここで、第１の値は正の値とする。その後、学習部３３０は、処理をＳ１９に移す。

Ｓ１７において、報酬算出部３３１は、報酬を第２の値とする。ここで、第２の値は正の値とする。また、第２の値は、第１の値より小さいものとする。その後、学習部３３０は、処理をＳ１９に移す。

Ｓ１８において、報酬算出部３３１は、報酬を第３の値とする。ここで、第３の値は負の値とする。

なお、第１の値、第２の値及び第３の値については、前回と比較した時間の差の大きさによって、更に重みづけを行うようにしてもよい。

【００７４】

Ｓ１９において、価値関数更新部３３２は、上述にて算出された報酬の値に基づいて、価値関数記憶部３４０が記憶している価値関数Ｑを更新する。そして、学習部３３０は、再度Ｓ１１に戻り、上述した処理を繰り返すことにより、価値関数Ｑは、適切な値に収束していく。なお、学習部３３０は、上述した処理を、所定回数繰り返したことや、所定時間繰り返したことを条件として終了するようにしてもよい。

以上、行動情報学習装置３００の動作について説明したが、例えば、Ｓ１４からＳ１８にかけての報酬の値を算出する処理は、一例であって、これに限定されない。例えば、上述したように、状態ｓ´と状態ｓにおける受渡時間Ｔの偏差、移動時間Ｔ１の偏差、及び開放時間Ｔ２の偏差と、を予め設定された報酬対応テーブル（仮称）又は報酬関数（仮称）に入力して、報酬の値を算出するようにしてもよい。

【００７５】

以上、図４を参照して説明した動作により、本実施形態では、受渡時間Ｔ及び開放時間Ｔ２を短縮するための行動情報を生成するための価値関数Ｑを生成することができる、という効果を奏する。

【００７６】

次に、図５のフローチャートを参照して、行動情報学習装置３００による最適化行動情報の生成時の動作について説明をする。

まず、Ｓ２１において、行動情報学習装置３００の最適化行動情報出力部３５０は、価値関数記憶部３４０が記憶している価値関数Ｑを取得する。この価値関数Ｑは、上述したように価値関数更新部３３２がＱ学習を行うことにより更新したものである。

【００７７】

Ｓ２２において、最適化行動情報出力部３５０は、この価値関数Ｑに基づいて、例えば現在設定されている状態ｓにおいて、取り得る行動ａのなかから価値Ｑ（ｓ，ａ）の最も高い行動ａを最適な行動として選択することで最適化行動情報を生成し、生成した最適化行動情報をロボット制御装置２００に対して出力する。

【００７８】

以上により、ロボット制御装置２００が、この最適化行動情報に基づいて現在設定されている状態ｓ（すなわち、現在設定されているロボット制御プログラム及びパラメータ）を修正して、動作指令を生成する。そして、ロボット制御装置２００は、生成した動作指令をロボット１００に送ることにより、ロボット１００は、受渡時間Ｔが最短になるように動作することができる、という効果を奏する。

【００７９】

また、図５を参照して説明した動作により、本実施形態では、行動情報学習装置３００は、価値関数Ｑに基づいて最適化行動情報を生成し、ロボット制御装置２００は、この最適化行動情報に基づいて現在設定されているロボット制御プログラム及びパラメータを修正して、動作指令を生成する。そして、ロボット制御装置２００は、生成した動作指令をロボット１００に送ることにより、受渡時間Ｔを短縮して、ロボット１００を制御することが可能となる、という効果も奏する。

【００８０】

本実施形態では、上述したように、ロボット制御プログラムやパラメータの設定値を調整しながら強化学習を行うことにより、受渡時間Ｔを短縮することができる。すなわち、本実施形態は、一般的な技術に比べて、有利な効果を奏する。

【００８１】

なお、上記のロボット制御システム１０００に含まれる各装置のそれぞれは、ハードウェア、ソフトウェア又はこれらの組み合わせにより実現することができる。また、上記のロボット制御システム１０００に含まれる各装置のそれぞれの協働により行われる行動情報学習方法も、ハードウェア、ソフトウェア又はこれらの組み合わせにより実現することができる。ここで、ソフトウェアによって実現されるとは、コンピュータがプログラムを読み込んで実行することにより実現されることを意味する。

【００８２】

プログラムは、様々なタイプの非一時的なコンピュータ可読媒体（ｎｏｎ－ｔｒａｎｓｉｔｏｒｙ　ｃｏｍｐｕｔｅｒ　ｒｅａｄａｂｌｅ　ｍｅｄｉｕｍ）を用いて格納され、コンピュータに供給することができる。非一時的なコンピュータ可読媒体は、様々なタイプの実体のある記録媒体（ｔａｎｇｉｂｌｅ　ｓｔｏｒａｇｅ　ｍｅｄｉｕｍ）を含む。非一時的なコンピュータ可読媒体の例は、磁気記録媒体（例えば、フレキシブルディスク、磁気テープ、ハードディスクドライブ）、光磁気記録媒体（例えば、光磁気ディスク）、ＣＤ－ＲＯＭ（Ｒｅａｄ　Ｏｎｌｙ　Ｍｅｍｏｒｙ）、ＣＤ－Ｒ、ＣＤ－Ｒ／Ｗ、半導体メモリ（例えば、マスクＲＯＭ、ＰＲＯＭ（Ｐｒｏｇｒａｍｍａｂｌｅ　ＲＯＭ）、ＥＰＲＯＭ（Ｅｒａｓａｂｌｅ　ＰＲＯＭ）、フラッシュＲＯＭ、ＲＡＭ（ｒａｎｄｏｍ　ａｃｃｅｓｓ　ｍｅｍｏｒｙ））を含む。また、プログラムは、様々なタイプの一時的なコンピュータ可読媒体（ｔｒａｎｓｉｔｏｒｙ　ｃｏｍｐｕｔｅｒ　ｒｅａｄａｂｌｅ　ｍｅｄｉｕｍ）によってコンピュータに供給されてもよい。一時的なコンピュータ可読媒体の例は、電気信号、光信号、及び電磁波を含む。一時的なコンピュータ可読媒体は、電線及び光ファイバ等の有線通信路、又は無線通信路を介して、プログラムをコンピュータに供給できる。

【００８３】

また、上述した実施形態は、本発明の好適な実施形態ではあるが、上記実施形態のみに

本発明の範囲を限定するものではなく、本発明の要旨を逸脱しない範囲において種々の変更を施した形態での実施が可能である。

【００８４】

上述した実施形態では、行動情報学習装置３００を、ロボット１００やロボット制御装置２００とは別体の装置により実現することを想定していたが、行動情報学習装置３００の機能の一部又は全部を、例えば、ロボット制御装置２００により実現するようにしてもよい。

【００８５】

上述した実施形態では、行動情報学習装置３００を、学習を行う機能と、行動情報を生成する機能とを有するものとしたが、学習を行う機能と、行動情報を生成する機能とを別の装置で行うようにしてもよい。

【００８６】

上述した実施形態では、行動情報学習装置３００が強化学習を行うものを説明した。この点、図６に示すように、ロボット制御システム１０００－２が、ｍ個の行動情報学習装置３００に対してネットワーク５００を介して接続された管理装置６００を備えるものとしてもよい。例えば、ロボット制御装置２００ごとに行動情報学習装置３００を備えた場合には、ｍは、ロボット制御装置２００の数である。

そして、ロボット１００と作業者Ｐの相対的な作業環境が同じ条件（例えば、ロボット１００の位置、受渡領域８、ロボット１００のハンド部１３の移動可能領域等が相対的に同じであること）を満たす場合、複数の行動情報学習装置３００－１～３００－ｍに対してネットワーク５００を介して管理装置６００を接続することで、管理装置６００は、各行動情報学習装置３００の価値関数Ｑを集約することができる。そうすることで、価値関数Ｑは、すべての行動情報学習装置３００との間で共有される。価値関数Ｑを複数の行動情報学習装置３００で共有するようにすれば、各行動情報学習装置３００にて分散して強化学習を行うことが可能となるので、強化学習の効率を向上させることが可能となる。

【００８７】

そして、管理装置６００が、集約した価値関数Ｑを、各行動情報学習装置３００に対して送信するようにしてもよい。

なお、管理装置６００は、各行動情報学習装置３００から学習用のデータを収集し、価値関数Ｑを更新するようにしてもよい。

また、管理装置６００が、最適化行動情報を、各ロボット制御装置２００に対して出力するようにしてもよい。

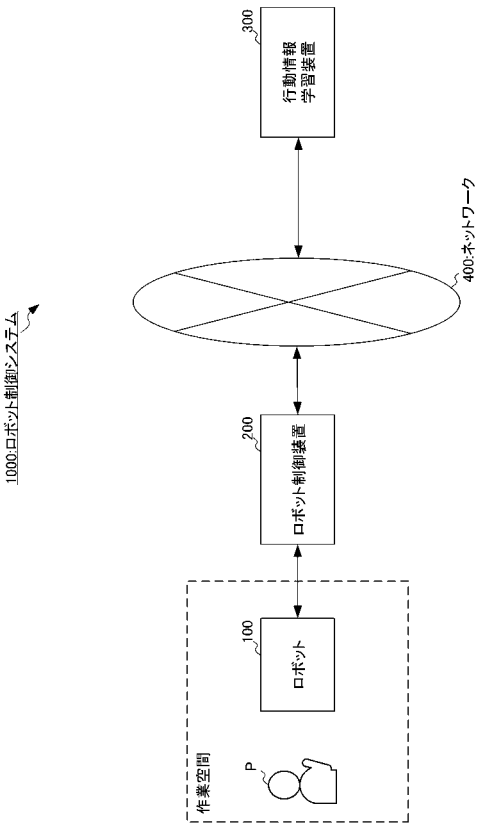
【符号の説明】

【００８８】

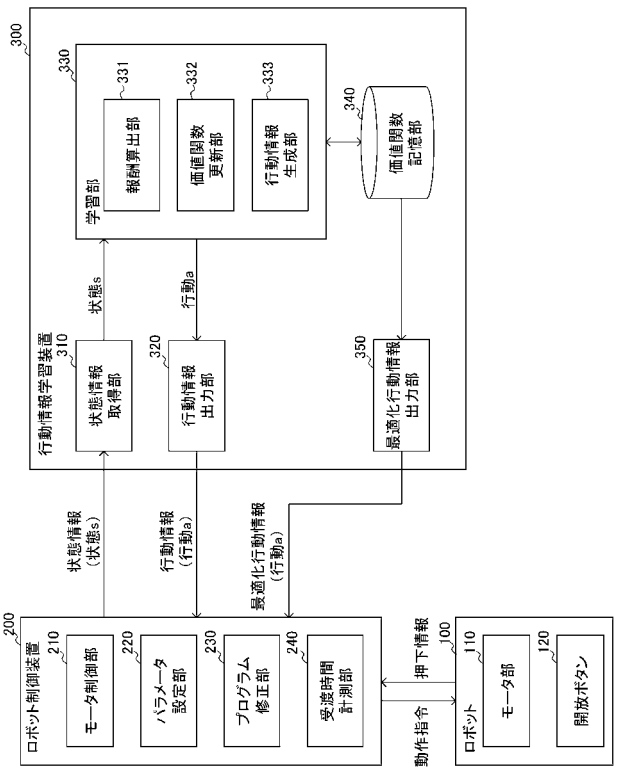
７　　ワーク
８　　受渡領域
１３　　ハンド部
１００　ロボット
１１０　モータ部
１２０　開放ボタン
２００　ロボット制御装置
２１０　モータ制御部
２２０　パラメータ設定部
３００　行動情報学習装置
３１０　状態情報取得部
３２０　行動情報出力部
３３０　学習部
３３１　報酬算出部
３３２　価値関数更新部
３３３　行動情報生成部

３４０　価値関数記憶部
３５０　最適化行動情報出力部
４００，５００　ネットワーク
１０００　ロボット制御システム
Ｐ　作業者
Ｔ　受渡時間
Ｔ１　移動時間
Ｔ２　開放時間

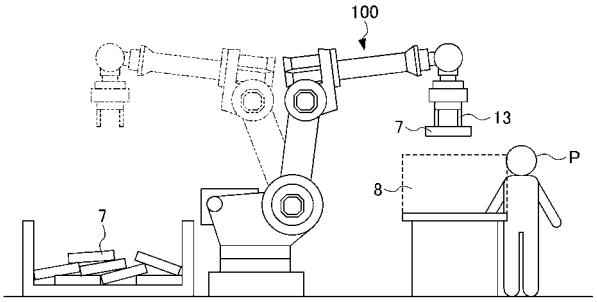
【図１】



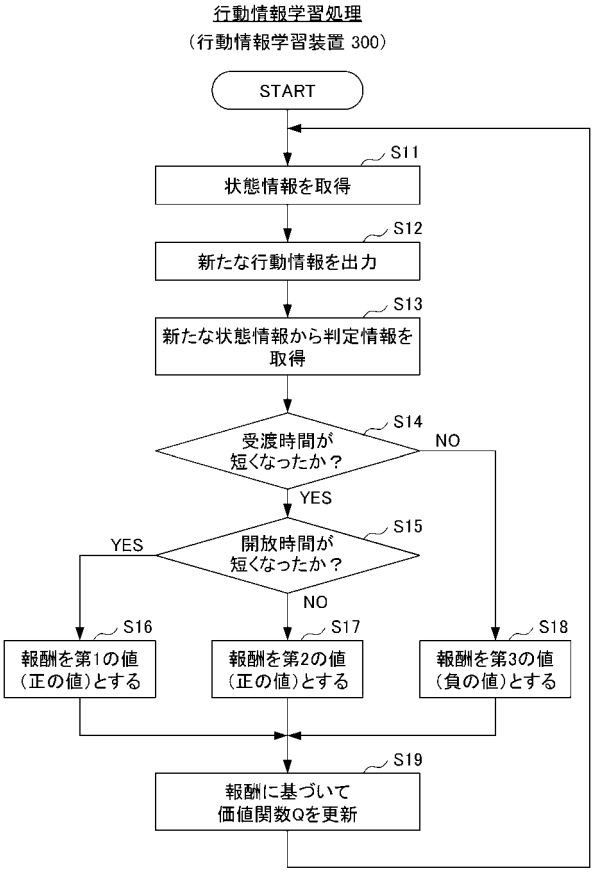

【図２】



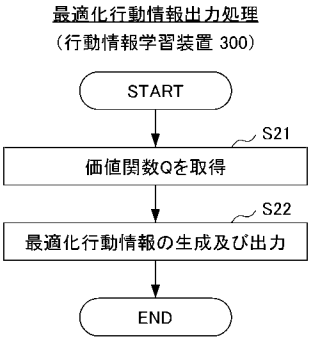
【図３】


100
7
13
P
8
7

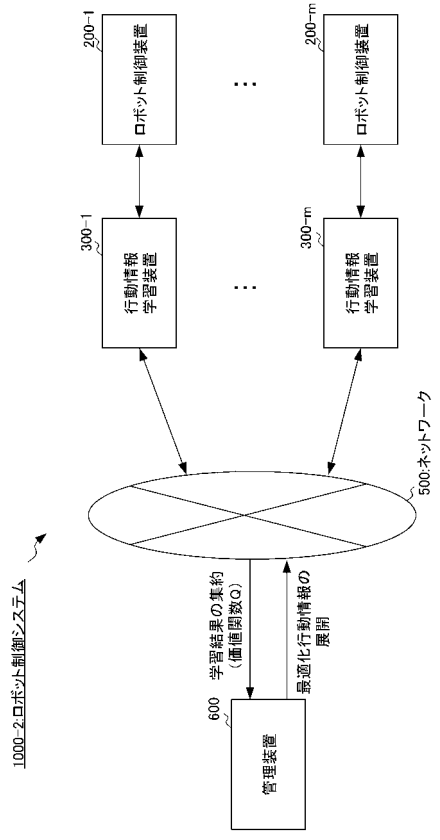
【図４】


行動情報学習処理
（行動情報学習装置 300）
START
S11
状態情報を取得
S12
新たな行動情報を出力
S13
新たな状態情報から判定情報を取得
S14
受渡時間が短くなったか？
NO
YES
S15
開放時間が短くなったか？
YES
NO
S16
報酬を第1の値（正の値）とする
S17
報酬を第2の値（正の値）とする
S18
報酬を第3の値（負の値）とする
S19
報酬に基づいて価値関数Qを更新

【図５】


最適化行動情報出力処理
（行動情報学習装置 300）
START
S21
価値関数Qを取得
S22
最適化行動情報の生成及び出力
END

【図６】


200-1
ロボット制御装置
200-m
ロボット制御装置
300-1
行動情報学習装置
300-m
行動情報学習装置
500:ネットワーク
1000-2:ロボット制御システム
学習結果の集約（価値関数Q）
最適化行動情報の展開
600
管理装置

フロントページの続き

(72)発明者　栗原　佑典
山梨県南都留郡忍野村忍草字古馬場３５８０番地　ファナック株式会社内

Ｆターム(参考) 3C269 AB33 BB08 EF60 MN27 MN44
3C707 AS01 BS12 KS38 LS14 LV01 LW08 LW12