(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7357537号
(P7357537)

(45)発行日　令和5年10月6日(2023.10.6)

(24)登録日　令和5年9月28日(2023.9.28)

(51)国際特許分類

Ｇ０５Ｂ　13/02　(2006.01)	Ｇ０５Ｂ	13/02	Ｌ
Ｇ０６Ｎ　20/00　(2019.01)	Ｇ０６Ｎ	20/00	
Ｇ０５Ｂ　23/02　(2006.01)	Ｇ０５Ｂ	23/02	３０１Ｚ

ＦＩ

請求項の数　16　（全21頁）

(21)出願番号　特願2019-233323(P2019-233323)
(22)出願日　令和1年12月24日(2019.12.24)
(65)公開番号　特開2021-103356(P2021-103356A)
(43)公開日　令和3年7月15日(2021.7.15)
審査請求日　令和3年11月26日(2021.11.26)

(73)特許権者　000005326
本田技研工業株式会社
東京都港区南青山二丁目１番１号
(74)代理人　110003281
弁理士法人大塚国際特許事務所
(74)代理人　100076428
弁理士　大塚 康徳
(74)代理人　100115071
弁理士　大塚 康弘
(74)代理人　100112508
弁理士　高柳 司郎
(74)代理人　100116894
弁理士　木村 秀二
(74)代理人　100134175
弁理士　永川 行光

最終頁に続く

(54)【発明の名称】　制御装置、制御装置の制御方法、プログラム、情報処理サーバ、情報処理方法、並びに制御システム

(57)【特許請求の範囲】

【請求項１】

強化学習を用いて車両に対する所定の制御を行う制御装置であって、

前記車両のライフサイクルにおけるイベントを検知する検知手段と、

前記イベントが検知されたことに応じて、検知された前記イベントに応じて特定される探索パラメータを、前記強化学習における探索の割合を調整する値として設定する設定手段と、

設定された前記探索パラメータに従って前記強化学習を用いた前記所定の制御を実行する処理手段と、を有し、

前記設定手段は、前記車両の利用開始までの前記車両に対する手続きの完了、又は前記車両を制御するための前記強化学習に用いられる学習モデルのバージョンの更新に関する第１のイベントが検知された場合、前記第１のイベントの後である第１の期間に設定される探索の割合を、前記第１のイベントが検知される前の第２の期間に設定されていた探索の割合より小さくする前記探索パラメータを設定する、ことを特徴とする制御装置。

【請求項２】

前記設定手段は、前記第１の期間に設定される探索の割合を非ゼロとする前記探索パラメータを設定する、ことを特徴とする請求項１に記載の制御装置。

【請求項３】

前記設定手段は、前記第１の期間に設定される探索の割合と、前記第２の期間に設定されていた探索の割合とが非連続となる前記探索パラメータを設定する、ことを特徴とする

請求項１または２に記載の制御装置。
【請求項４】
前記第１のイベントは、更に、前記車両の特定の使用状態への到達を含む、ことを特徴とする請求項１から３のいずれか１項に記載の制御装置。
【請求項５】
前記第１のイベントは、前記車両の利用開始までの前記車両に対する手続きの完了を含み、当該手続きの完了は、前記車両の製造完了、及び、前記車両の登録完了の少なくともいずれかを含む、ことを特徴とする請求項１に記載の制御装置。
【請求項６】
前記車両の特定の使用状態への到達は、所定の時点からの所定日数の経過、所定の時点からの所定走行距離の走行の少なくともいずれかを含む、ことを特徴とする請求項４に記載の制御装置。
【請求項７】
前記第１のイベントは、前記車両を制御するための前記強化学習に用いられる学習モデルのバージョンの更新を含む、ことを特徴とする請求項１に記載の制御装置。
【請求項８】
検知された前記イベントに応じて、前記探索パラメータを特定する特定手段を更に有する、ことを特徴とする請求項１から７のいずれか１項に記載の制御装置。
【請求項９】
検知された前記イベントを外部サーバに送信する送信手段と、
前記イベントに応じて特定された前記探索パラメータを前記外部サーバから受信する受信手段と、を更に有する、ことを特徴とする請求項１から７のいずれか１項に記載の制御装置。
【請求項１０】
前記探索パラメータは、車両ごと、又は車両のモデルごとに異なる、ことを特徴とする、請求項１から９のいずれか１項に記載の制御装置。
【請求項１１】
前記処理手段によって実行される前記強化学習のモデルに対する入力情報と出力情報とを、学習データとして外部サーバに提供する提供手段を更に有する、ことを特徴とする請求項１から１０のいずれか１項に記載の制御装置。
【請求項１２】
強化学習を用いて車両に対する所定の制御を行う制御装置の制御方法であって、
検知手段が、前記車両のライフサイクルにおけるイベントを検知する検知工程と、
設定手段が、前記イベントが検知されたことに応じて、検知された前記イベントに応じて特定される探索パラメータを、前記強化学習における探索の割合を調整する値として設定する設定工程と、
処理手段が、設定された前記探索パラメータに従って前記強化学習を用いた前記所定の制御を実行する処理工程と、を有し、
前記設定工程では、前記車両の利用開始までの前記車両に対する手続きの完了、又は前記車両を制御するための前記強化学習に用いられる学習モデルのバージョンの更新に関する第１のイベントが検知された場合、前記第１のイベントの後である第１の期間に設定される探索の割合を、前記第１のイベントが検知される前の第２の期間に設定されていた探索の割合より小さくする前記探索パラメータを設定する、ことを特徴とする制御装置の制御方法。
【請求項１３】
コンピュータを、請求項１から１１のいずれか１項に記載の制御装置の各手段として機能させるためのプログラム。
【請求項１４】
強化学習を用いて車両に対する所定の制御を行う情報処理サーバであって、
前記車両のライフサイクルにおけるイベントを検知する検知手段と、

前記イベントが検知されたことに応じて、検知された前記イベントに応じて特定される探索パラメータを、前記強化学習における探索の割合を調整する値として設定する設定手段と、
設定された前記探索パラメータに従って前記強化学習を用いた前記所定の制御のための処理を実行する処理手段と、
前記処理手段による処理結果を前記車両に送信する送信手段と、を有し、
前記設定手段は、前記車両の利用開始までの前記車両に対する手続きの完了、又は前記車両を制御するための前記強化学習に用いられる学習モデルのバージョンの更新に関する第１のイベントが検知された場合、前記第１のイベントの後である第１の期間に設定される探索の割合を、前記第１のイベントが検知される前の第２の期間に設定されていた探索の割合より小さくする前記探索パラメータを設定する、ことを特徴とする情報処理サーバ。
【請求項１５】
情報処理サーバで実行される、強化学習を用いて車両に対する所定の制御を行う情報処理方法であって、
検知手段が、前記車両のライフサイクルにおけるイベントを検知する検知工程と、
設定手段が、前記イベントが検知されたことに応じて、検知された前記イベントに応じて特定される探索パラメータを、前記強化学習における探索の割合を調整する値として設定する設定工程と、
処理手段が、設定された前記探索パラメータに従って前記強化学習を用いた前記所定の制御のための処理を実行する処理工程と、
送信手段が、処理工程における処理結果を前記車両に送信する送信工程と、を有し、
前記設定工程では、前記車両の利用開始までの前記車両に対する手続きの完了、又は前記車両を制御するための前記強化学習に用いられる学習モデルのバージョンの更新に関する第１のイベントが検知された場合、前記第１のイベントの後である第１の期間に設定される探索の割合を、前記第１のイベントが検知される前の第２の期間に設定されていた探索の割合より小さくする前記探索パラメータを設定する、ことを特徴とする情報処理方法。
【請求項１６】
強化学習を用いて車両に対する所定の制御を行う制御装置と、情報処理サーバとを含む制御システムであって、
前記制御装置は、
前記車両のライフサイクルにおけるイベントを検知する検知手段と、
前記イベントが検知されたことに応じて、検知された前記イベントを前記情報処理サーバに送信する第１の送信手段と、
前記情報処理サーバから受信した、前記イベントに応じて特定された探索パラメータを、前記強化学習における探索の割合を調整する値として設定する設定手段と、
設定された前記探索パラメータに従って前記強化学習を用いた前記所定の制御を実行する処理手段と、を有し、
前記設定手段は、前記車両の利用開始までの前記車両に対する手続きの完了、又は前記車両を制御するための前記強化学習に用いられる学習モデルのバージョンの更新に関する第１のイベントが検知された場合、前記第１のイベントの後である第１の期間に設定される探索の割合を、前記第１のイベントが検知される前の第２の期間に設定されていた探索の割合より小さくする前記探索パラメータを設定する、制御装置と、
前記情報処理サーバは、
前記イベントに応じて、前記探索パラメータを特定する特定手段と、
特定した前記探索パラメータを前記車両に送信する第２の送信手段と、を有する、ことを特徴とする制御システム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、制御装置、制御装置の制御方法、プログラム、情報処理サーバ、情報処理方

法、並びに制御システムに関する。
【背景技術】
【０００２】
近年、車両の自律走行を実現するための行動制御に強化学習を用いる技術が知られている（特許文献１）。特許文献１には、強化学習における方策（行動選択ルールをいう。ポリシーともいわれる）を学習する過程で、確率εでランダムに行動を選択し、確率１－εで方策に従って行動を選択する（ε－ｇｒｅｅｄｙ法ともいわれる）ことが開示されている。すなわち、より適切な方策を学習によって獲得するためには、より多様な行動方策を得るための探索（ｅｘｐｌｏｒａｔｉｏｎ）と学習した方策の活用（ｅｘｐｌｏｉｔａｔｉｏｎ）の両立が必要となる。
【先行技術文献】
【特許文献】
【０００３】
【文献】特願２０１９－０８７０９６号公報
【発明の概要】
【発明が解決しようとする課題】
【０００４】
ところで、強化学習によって行動制御を行う商品化された車両が、市場を走行する場合、学習済みの状態でテスト等がなされた一意な制御信号が出力されるように、行動制御における探索を行わないことが想定される。一方、自律走行のような高次元の行動制御を実現するためには、広大な行動空間から最適な行動を探索するための学習が必要であり、そのための学習データ、とりわけ実環境で得られる様々なデータを、車両の市場導入後も収集し、活用することが望まれる場合がある。
【０００５】
本発明は、上記課題に鑑みてなされ、その目的は、車両の制御において、車両のライフサイクルにおいて強化学習における探索と活用を継続的に両立することが可能な技術を提供することである。
【課題を解決するための手段】
【０００６】
本発明によれば、
強化学習を用いて車両に対する所定の制御を行う制御装置であって、
前記車両のライフサイクルにおけるイベントを検知する検知手段と、
前記イベントが検知されたことに応じて、検知された前記イベントに応じて特定される探索パラメータを、前記強化学習における探索の割合を調整する値として設定する設定手段と、
設定された前記探索パラメータに従って前記強化学習を用いた前記所定の制御を実行する処理手段と、を有し、
前記設定手段は、前記車両の利用開始までの前記車両に対する手続きの完了、又は前記車両を制御するための前記強化学習に用いられる学習モデルのバージョンの更新に関する第１のイベントが検知された場合、前記第１のイベントの後である第１の期間に設定される探索の割合を、前記第１のイベントが検知される前の第２の期間に設定されていた探索の割合より小さくする前記探索パラメータを設定する、ことを特徴とする制御装置が提供される。
【発明の効果】
【０００７】
本発明によれば、車両の制御において、車両のライフサイクルにおいて強化学習における探索と活用を継続的に両立することが可能になる。
【図面の簡単な説明】
【０００８】
【図１】本発明の実施形態に係る車両制御システムの概要を示す図

【図２】本実施形態に係る車両の機能構成例を示すブロック図
【図３】本実施形態に係る強化学習を用いた制御の一例としてのダンパ制御の動作概要を説明する図
【図４】本実施形態に係るモデル処理部における強化学習の一例として、アクタークリティック手法を適用する場合の構成を説明する図
【図５】本実施形態において利用可能なセンサ及び当該センサにより計測されるセンサデータの例を示す図
【図６】本実施形態に係る探索パラメータの変化の一例を示す図
【図７】本実施形態に係る車両におけるダンパ制御処理の一連の動作を示すフローチャート
【図８】本実施形態に係る車両における探索パラメータ設定処理の一連の動作を示すフローチャート
【図９】本実施形態に係る情報処理サーバの一例としてのデータ収集サーバの機能構成例を示す図
【図１０】本実施形態に係るデータ収集サーバにおける探索パラメータ送信処理の一連の動作を示すフローチャート
【図１１】本実施形態に係るイベントと探索パラメータの値との対応付けを説明するための図
【発明を実施するための形態】
【０００９】
以下、添付図面を参照して実施形態を詳しく説明する。尚、以下の実施形態は特許請求の範囲に係る発明を限定するものでするものでなく、また実施形態で説明されている特徴の組み合わせの全てが発明に必須のものとは限らない。実施形態で説明されている複数の特徴うち二つ以上の特徴が任意に組み合わされてもよい。また、同一若しくは同様の構成には同一の参照番号を付し、重複した説明は省略する。
【００１０】
＜車両制御システムの概要＞
図１を参照して、本実施形態に係る車両制御システム１０の概要について説明する。車両制御システム１０は、所定システムの一例としての車両１００、および情報処理サーバの一例としてのデータ収集サーバ１１０とを含む。本実施形態では、車両１００が、強化学習を用いて車両の構成要素であるダンパを制御する場合を例に説明する。しかし、車両が強化学習を用いて、ダンパ制御以外の他の構成要素を制御（例えば走行時の操舵やアクセル制御）を行う場合にも適用可能である。また、以下で説明する実施形態では、車両が備える制御部が強化学習を用いたダンパ制御を行う場合を例に説明する。しかし、制御部による処理を、車両内に搭載された情報処理装置が行うようにしてもよい。すなわち、本実施形態は、情報処理装置が、車両からセンサ情報等を取得して、強化学習を用いたダンパ制御用の制御信号を車両に出力する構成によって実現されてもよい。なお、以下の説明では、車両が備える制御部或いは上記情報処理装置を制御装置という場合がある。
【００１１】
データ収集サーバ１１０は、強化学習を用いた学習モデルを学習させるための学習データを取得して蓄積するためのサーバである。データ収集サーバ１１０は、実環境において制御が行われている複数の車両１００で収集される学習データを、それぞれの車両１００から取得する。学習データは、詳細は後述するが、車両１００のセンサで取得されるフィードバックデータを含む。学習データは、報酬や行動などの他の情報（すなわち強化学習で用いる入出力の情報）を含んでよい。データ収集サーバ１１０は、車両１００から取得した学習データをデータベース（ＤＢ）１１１に蓄積する。
【００１２】
本実施形態のデータ収集サーバ１１０は、実環境において制御が行われている車両１００からイベント情報を取得し、取得したイベント情報に応じて車両１００のパラメータ制御を行う。イベント情報は、車両１００のライフサイクルにおいて発生するイベントを示す情報である。イベントは、例えば、車両に対する手続きの完了（車両製造完了、車両登

録の完了）や車両の特定の使用状態への到達（製造完了から所定日数の経過、製造完了から所定走行距離の走行）、或いは、車両を制御する構成要素の更新（学習モデルのバージョンの所定回数の更新実施）などを含む。

【００１３】

車両１００は、強化学習を用いた制御において方策を決定する際に、例えば、所定の確率εでランダムな行動を選択（すなわち探索）し、１－εの確率で方策の出力に従って行動を選択（すなわち活用）する。このような車両１００に対し、データ収集サーバ１１０は、それぞれの車両１００のライフサイクルに応じたイベントの発生に応じて、車両ごとのパラメータ（ここでは確率ε）を制御する。車両１００は、データ収集サーバ１１０から指定されたパラメータ（確率ε）に従って、強化学習における探索と活用のバランスを両立させる。このようにすることで、データ収集サーバ１１０は、ある程度のばらつきを持った多様なデータを実環境における車両１００から収集することができる。ひいては、収集した学習データを用いたモデルの性能をより高性能化することが可能になる。なお、後述するように、強化学習アルゴリズムの探索によって安全上許容できない出力が選択される場合、当該出力によって車両が制御されないように出力値はフィルタアウトされる。

【００１４】

＜車両の構成＞

次に、図２を参照して、本実施形態に係る車両１００の機能構成例について説明する。なお、以降の図を参照して説明する機能ブロックの各々は、統合されまたは分離されてもよく、また説明する機能が別のブロックで実現されてもよい。また、ハードウェアとして説明するものがソフトウェアで実現されてもよく、その逆であってもよい。

【００１５】

センサ部１０１は、車両１００に備えられる各種センサであり、車両１００の挙動に関するセンサデータを出力する。図５は、センサ部１０１のうち、本実施形態のダンパ制御処理に用いられ得る各種センサと計測内容の一例を示している。これらのセンサには、例えば、車両１００の車速を計測するための車速センサや、車両のボディ加速度を計測するための加速度センサ、ダンパのストローク挙動（速度や変位）を計測するサスペンション変位センサを含む。更に、ステアリング入力を計測する操舵角センサ、自己位置を取得するＧＰＳ等が含まれる。なお、以降の説明では、ダンパ制御処理に用いられる、車両１００の挙動に関するこれらのセンサデータを特にフィードバックデータという。センサ部１０１から出力された車両１００の挙動に関するフィードバックデータは、制御部２００やダンパ制御部１０６に入力される。

【００１６】

また、センサ部１０１は、車両の前方（或いは、更に後方方向や周囲）を撮影した撮影画像を出力する撮影用カメラや、車両の前方（或いは、更に後方方向や周囲）の距離を計測して得られる距離画像を出力するＬｉｄａｒ（Ｌｉｇｈｔ Ｄｅｔｅｃｔｉｏｎ ａｎｄ Ｒａｎｇｉｎｇ）を含んでよい。例えば、制御部２００は、撮影画像や距離画像など空間的な情報をフィードバックデータとして強化学習を用いたダンパ制御或いは車両１００の行動制御を行ってもよい。

【００１７】

通信部１０２は、例えば通信用回路等を含む通信デバイスであり、例えばＬＴＥやＬＴＥ－Ａｄｖａｎｃｅｄ等或いは所謂５Ｇとして規格化された移動体通信を介して外部のサーバや周囲の交通システムなどと通信する。地図データの一部又は全部を外部サーバから受信したり、他の交通システムから交通情報などを受信し得る。また、通信部１０２は、センサ部１０１から取得された各種データ（例えば、フィードバックデータ）やイベント情報をデータ収集サーバ１１０に送信する。そして、通信部１０２は、データ収集サーバ１１０から、パラメータ制御に係る情報（例えば探索を行うための確率εを示す情報。以下、探索パラメータという）を受信する。

【００１８】

操作部１０３は、車両１００内に取り付けられたボタンやタッチパネルなどの操作部材

のほか、ステアリングやブレーキペダルなどの、車両１００を運転するための入力を受け付ける部材を含む。電源部１０４は、例えばリチウムイオンバッテリ等で構成されるバッテリを含み、車両１００内の各部に電力を供給する。動力部１０５は、例えば車両を走行させるための動力を発生させるエンジンやモータを含む。

【００１９】

ダンパ１０７は、車両１００のサスペンションに用いられ、例えば、ダンパの特性である減衰力を制御可能なアクティブダンパである。例えば、ダンパ１０７の制御は、ダンパ１０７の内部のコイルに流す電流量を制御することで内部のバルブが開く圧力が調節され、ダンパ１０７の減衰力が制御される。ダンパ１０７は、それぞれ独立する４つのダンパ１０７で構成され、それぞれ独立して制御される。なお、車両１００が強化学習を用いて（ダンパ制御とは異なる制御である）車両の行動制御などを行う場合、ダンパ１０７は通常のダンパであってもかまわない。

【００２０】

ダンパ制御部１０６は、ダンパ１０７の特性を制御するための例えばソフトウェアモジュールであり、ダンパ制御部１０６は、制御部２００から出力される制御変数に基づいて（独立した４つのダンパ１０７のそれぞれの）ダンパの特性を制御する。なお、本実施形態では、ダンパ制御に求められる高速応答性を確保するために、ダンパ制御部１０６がダンパ１０７を制御するようにしているが、ダンパ制御部１０６は必ずしも必須ではなく、制御部２００がダンパ１０７を直接制御するようにしてもよい。

【００２１】

記憶部１０８は、半導体メモリなどの不揮発性の大容量のストレージデバイスを含む。センサ部１０１から出力されたフィードバックデータ、或いは、制御部２００によって選別されたフィードバックデータを、データ収集サーバ１１０に送信するために、一時的に格納する。

【００２２】

制御部２００は、例えば、ＣＰＵ２１０、ＲＡＭ２１１、ＲＯＭ２１２を含み、車両１００の各部の動作を制御する。また、制御部２００は、センサ部１０１からフィードバックデータを取得して、ダンパ制御処理を実行したり、データ収集サーバ１１０から受信した探索パラメータに応じて、強化学習における探索と活用のバランスを制御する。制御部２００は、ＣＰＵ２１０がＲＯＭ２１２に格納されたコンピュータプログラムを、ＲＡＭ２１１に展開、実行することにより、データ入力部２１３、モデル処理部２１４、報酬決定部２１５、探索パラメータ設定部２１６の機能を発揮させる。

【００２３】

ＣＰＵ２１０は、１つ以上のプロセッサを含む。ＲＡＭ２１１は、例えばＤＲＡＭ等を含み、ＣＰＵ２１０のワークメモリとして機能する。ＲＯＭ２１２は、不揮発性の記憶媒体で構成され、ＣＰＵ２１０によって実行されるコンピュータプログラムや制御部２００を動作させる際の設定値などを記憶する。なお、以下の実施形態では、ＣＰＵ２１０がモデル処理部２１４の処理を実行する場合を例に説明するが、モデル処理部２１４の処理は不図示の１つ以上の他のプロセッサ（例えばＧＰＵ）で実行されてもよい。

【００２４】

データ入力部２１３は、記憶部１０８に記憶されたフィードバックデータを取得して、データの前処理を行う。フィードバックデータとして入力される車両の運動状態や運転入力の特徴を、機械学習アルゴリズムが処理し易いように、種々の加工処理を行う。加工処理の一例では、所定の期間内のフィードバックデータの最大値、最小値等に加工する処理を含む。事前にフィードバックデータを加工しておくことにより、生のフィードバックデータを機械学習アルゴリズムで直接扱う場合よりも処理効率や学習効率を向上させることができる。なお、本実施形態の例では、データ入力部２１３によって加工したフィードバックデータを、学習データとして、データ収集サーバ１１０に送信する場合を例に説明する。しかし、データ入力部２１３による加工を行っていない状態のフィードバックデータを、学習データとして強化学習に用いたり、データ収集サーバ１１０に送信したりしても

よい。

【００２５】

モデル処理部２１４は、例えば、深層強化学習などの機械学習アルゴリズムの演算を行って、得られた出力をダンパ制御部１０６に出力する。モデル処理部２１４は、データ入力部２１３からのフィードバックデータと報酬決定部２１５からの報酬のデータを用いて、強化学習アルゴリズムを実行し、ダンパ制御部１０６に提供する制御変数を出力する。モデル処理部２１４は、強化学習アルゴリズムの実行を通して内部のパラメータを最適化し（すなわち学習し）、内部のパラメータで特定される演算処理をフィードバックデータに対して適用することにより、車両１００の挙動に応じた最適な制御変数を出力する。また、モデル処理部２１４は、方策に係るニューラルネットワーク（アクター）から出力される複数の行動から、探索パラメータに従って行動を選択する処理を含む。

【００２６】

報酬決定部２１５は、フィードバックデータに基づいて、機械学習アルゴリズム（強化学習アルゴリズム）で用いられる報酬又はペナルティを決定し、モデル処理部２１４に出力する。探索パラメータ設定部２１６は、データ収集サーバ１１０から取得した探索パラメータをモデル処理部２１４に設定する。

【００２７】

イベント検知部２１７は、車両１０１のセンサ部１０１によって計測された情報やモデル処理部２１４で動作する学習モデルのバージョン情報等に基づいて、車両１００のライフサイクルにおいて発生するイベントを検知し、検知したイベントをイベント情報としてデータ収集サーバ１１０に送信する。イベント情報は、車両１００のライフサイクルにおいて発生するイベントを示す情報である。イベントは、上述したように、例えば、車両に対する手続きの完了（車両製造完了、車両登録の完了）や車両の特定の使用状態への到達（製造完了から所定日数の経過、製造完了から所定走行距離の走行）、或いは、車両を制御する構成要素の更新（学習モデルのバージョンの所定回数の更新実施）などを含む。

【００２８】

＜強化学習を用いたダンパ制御処理の概要＞

次に、図３を参照して、強化学習を用いたダンパ制御処理の概要について説明する。

【００２９】

本実施形態のダンパ制御処理は、例えば、モデル処理部２１４における深層強化学習アルゴリズムを用いた演算処理と、ダンパ制御部１０６における演算処理とを含む。このような構成では、ダンパ制御部１０６は、予め決められたルールベースの演算処理により、低次元制御出力を数百ヘルツの高速な動作周波数でダンパを制御することができる一方、モデル処理部２１４はダンパ制御部ほど高くない動作周波数で高次元の制御を実行することができる。もちろん、ダンパ制御の構成は、この構成に限定されるものではなく、ダンパ制御部１０６を設けること無く、モデル処理部２１４が直接的にダンパ１０７の制御を行うようにしてもよい。

【００３０】

例えば、モデル処理部２１４は、ある時刻ｔにおいて、データ入力部２１３からのフィードバックデータを受け付けて強化学習アルゴリズムを実行し、得られた制御変数をダンパ制御部１０６に出力する。強化学習では、このフィードバックデータは環境の状態（s_t）に相当し、制御変数は、環境に対する行動（a_t）に相当する。

【００３１】

ダンパ制御部１０６は、モデル処理部２１４からの制御変数を受け付けると、ダンパ制御部１０６の内部で用いられている制御変数を、モデル処理部２１４から取得した新たな制御変数に置き換える。制御変数は、例えば、フィードバックデータに応じたゲインパラメータなどの、ダンパ制御部１０６がダンパの特性を決定するためのパラメータを含む。また、制御変数は、ダンパ制御部１０６が公知のスカイフック理論に基づいてダンパ１０７の減衰力を決定するためのパラメータでもある。例えば、車両１００のセンサ部１０１において計測される車両のボディ加速度がスカイフック理論に基づく加速度と整合するよ

うにダンパ１０７の減衰力が制御される。

【００３２】

ダンパ制御部１０６は、モデル処理部２１４からの新たな制御変数に基づいて、フィードバックデータに対するダンパ特性の制御を行う。このとき、ダンパ制御部１０６は、ダンパ１０７の特性を制御するための制御量を算出する。例えば、ダンパ１０７の特性は減衰力であり、ダンパ１０７の特性を制御するための制御量は、当該減衰力を制御する電流量である。ダンパ制御部１０６は、時刻がｔ＋１になるまで、新たな制御変数に基づく、フィードバックデータに対するダンパ制御を繰り返す。

【００３３】

センサ部１０１は、時刻ｔ＋１におけるフィードバックデータを取得し、データ入力部２１３は、このフィードバックデータを加工して、加工したフィードバックデータをモデル処理部２１４に出力する。強化学習では、この加工したフィードバックデータは、環境における状態（s_{t+1}）に相当する。報酬決定部２１５は、当該フィードバックデータに基づいて、強化学習における報酬（r_{t+1}）（またはペナルティ）を決定してモデル処理部２１４に提供する。本実施形態では、報酬は、所定のフィードバックデータの組み合わせから得られる、車両の挙動に関する報酬値である。

【００３４】

モデル処理部２１４は、報酬（r_{t+1}）を受け付けると、後述する方策および状態価値関数を更新して、時刻ｔ＋１におけるフィードバックデータに対する新たな制御変数を出力する（行動（a_{t+1}））。

【００３５】

＜モデル処理部の構成＞

更に、図４を参照して、モデル処理部２１４の構成例とダンパ制御処理におけるモデル処理部２１４の動作例について説明する。図４は、アクタークリティック手法を用いる場合のモデル処理部２１４の内部構成例と、モデル処理部２１４のニューラルネットワーク（ＮＮ）のネットワーク構成例を模式的に示している。

【００３６】

モデル処理部２１４は、アクター４０１とクリティック４０２とを含む。アクター４０１は、方策π（ｓ，ａ）に基づき行動（ａ）を選択する機構である。一例として、状態ｓで行動ａを選択する確率をｐ（ｓ，ａ）とすると、方策は、ｐ（ｓ，ａ）と例えばｓｏｆｔｍａｘ関数などを用いた所定の関数とで定義される。クリティック４０２は、現在アクターが利用している方策π（ｓ，ａ）に対する評価を行う機構であり、当該評価を表す状態価値関数Ｖ（ｓ）を有する。

【００３７】

図３において説明した時刻ｔから時刻ｔ＋１における動作を例に説明すると、ある時刻ｔにおいて、アクター４０１はフィードバックデータを受け付け、方策π（ｓ，ａ）に基づき制御変数（すなわち行動（a_t））を出力する。

【００３８】

ダンパ制御部１０６により、時刻ｔに対する制御変数を用いてダンパ制御が行われた後に、時刻ｔ＋１におけるフィードバックデータ（すなわち状態（s_{t+1}）が得られると、報酬決定部２１５から当該フィードバックデータに基づく報酬（r_{t+1}）がクリティック４０２に入力される。

【００３９】

クリティック４０２は、アクターの方策を改善するための方策改善を算出して、アクター４０１に入力する。方策改善は、公知の所定の計算方法によって求めたものでよいが、例えば、報酬とフィードバックデータを用いて得られる、公知のＴＤ誤差$\delta_t = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$（γは強化学習における割引報酬）を方策改善として用いることができる。

【００４０】

アクター４０１は、方策改善に基づいて方策π（ｓ，ａ）を更新する。方策の更新は、

例えば、p（s_t，a_t）をp（s_t，a_t）＋βδ_t（βはステップサイズパラメータ）で置き換えるような更新を行いうる。すなわち、アクター４０１は報酬に基づく方策改善を用いて方策を更新する。クリティック４０２は、状態価値関数Ｖ（ｓ）を、例えばＶ（ｓ）＋αδ_t（αはステップサイズパラメータ）で置き換えて更新する。

【００４１】

図４の右図は、モデル処理部２１４が用いる学習モデルをディープニューラルネットワーク（単にＮＮともいう）において実現する場合のネットワーク構成例を模式的に示している。この例では、アクターとクリティックの２つのニューラルネットワークで構成される。入力層４１０は、例えば１４５０個のニューロンで構成され、対応するフィードバックデータが入力される。

【００４２】

入力層４１０から入力された信号はそれぞれアクターの隠れ層４１１、クリティックの隠れ層４１２を順方向に伝搬してそれぞれの出力層４１３と４１４から出力値が得られる。アクターのＮＮからの出力は方策（取り得る行動）であり、クリティックのＮＮからの出力は状態価値である。一例として、アクターの隠れ層４１１は、例えば、５層のネットワーク構造で構成され、クリティックの隠れ層４１２は、例えば、３層のネットワーク構造で構成される。

【００４３】

アクターの出力層４１３は、例えば、２２個のニューロンで構成され、クリティックの出力層４１４は、例えば、１個のニューロンで構成される。例えば、出力層４１３のニューロンの列はとりうる行動のリストに対応付けられており、各ニューロンが、行動をとるべきスコア或いは行動のとられる確率を表してよい。出力層４１３において各ニューロンの値が出力されると、これらの複数の行動のなかから、探索パラメータに応じて行動が選択される。例えば探索パラメータが確立εである場合、確率εでランダムに行動を選択し、確率１－εで最も高いスコアを示す行動を選択する。なお、ネットワークのニューロン数や層の数、ネットワーク構成は適宜変更することができ、他の構成を用いてもよい。

【００４４】

それぞれのニューラルネットワークを最適化するためにニューラルネットワークの重みパラメータを変更する必要がある。ニューラルネットワークの重みパラメータの変更は、例えば、予め定めた損失関数を用いて誤差逆伝搬により行われる。本実施形態では、アクターとクリティックの２つのネットワークが存在するため、予めアクターの損失関数L_{actor}とクリティックの損失関数L_{critic}をそれぞれ用いる。それぞれのネットワークの重み付けパラメータは、例えば、各損失関数に対して所定の勾配降下方最適化手法（例えばＲＭＳｐｒｏｐ　ＳＧＤ）を用いることにより変更される。

【００４５】

制御部２００は、フィードバックデータ（状態s_t）を学習データとしてデータ収集サーバ１１０に送信する。あるいは、制御部２００は、当該フィードバックデータ（状態s_t）と対応するアクターの出力（行動a_t）と、報酬r_{t+1}と、行動a_tの結果生じたフィードバックデータ（状態s_{t+1}）とを１セットの学習データとして、データ収集サーバ１１０に送信してもよい。この場合、以下の説明において、単にフィードバックデータを学習データとして送信する旨の説明は、当該１セットの情報を学習データとして送信することを意味するものとして読み替えてよい。

【００４６】

＜イベントに応じた探索パラメータ設定処理の概要＞

次に、図６を参照して、車両１００のライフサイクルにおいて発生するイベントに応じてモデル処理部２１４に設定される探索パラメータの変化について説明する。

【００４７】

図６は、探索パラメータの値（縦軸）と時間（横軸）の関係を示しており、イベントが発生するごとに、探索パラメータの値が変化する様子を模式的に示している。探索パラメータは、強化学習アルゴリズムが確率εでランダムに行動を選択し、確率１－εで方策に

従って行動を選択する場合の確率εの値に対応する。また、時間は車両のライフサイクルに係る時間を表す。

【００４８】

イベント１の発生が、例えば、車両の製造完了時であるとする。この場合、車両の製造完了時より前（例えば開発時）から車両のライフサイクルを定義し、この期間にモデル処理部２１４の学習モデルが強化学習を行うことを想定している。この場合、車両の製造完了時より前の時間については、車両が実際に走行する場合のほか、例えばサーバ上でシミュレーション等により強化学習を進めているような場合であってもよい。もちろん、時間の原点を車両の製造完了時として、その後のイベントとして、イベント１、イベント２・・・が発生するものとしてもよい。

【００４９】

イベント１が発生するまでは、探索パラメータの値は値６０１となるように設定され、イベント１が発生したことに応じて、値６０２に設定される。イベント１が車両の製造完了である場合、このイベントが発生した後に設定される探索パラメータの値６０２は、イベント発生前に設定されている探索パラメータの値６０１より低い。そして、イベント発生前の探索パラメータと、イベント発生後の探索パラメータとが非連続となる探索パラメータが設定される。これは、イベント１が発生するまでに学習が進み、学習モデルの精度が向上していると考えられることから、イベント発生を契機として強化学習の探索的な要素を一段階下げることを意味する。但し、製造完了後も引き続き学習データの収集において探索的な要素を残し、ばらつきを含んだ学習データを収集するため、探索パラメータは０には設定しない。

【００５０】

同様に、順に、イベント２とイベント３が発生すると、その度に探索パラメータの値が引き下げられ、最終的にはｔ３以降では探索パラメータの値が０に設定されてもかまわない。イベント２やイベント３は、上述したように、例えば、車両１００の走行が製造完了から所定走行距離に達した場合や、学習モデルのバージョンが所定回数だけ更新された場合である。図６に示す例では、イベント３は、十分に学習モデルの精度が向上していると判定されるイベントに相当する。

【００５１】

なお、上記説明では、探索パラメータが値６０１～６０３のように、一定の値をとる場合を例に説明した。しかし、曲線６０４～６０６が示すように、２つのイベントの間の探索パラメータは、時刻経過、収集した学習データの量、或いは走行処理などの値に応じた関数の値として変化するものでもよい。この場合、曲線６０４～６０６に示す探索パラメータの値は、イベントの発生時において不連続となるように変化する。このようにすれば、例えば、イベント１とイベント２との間が長い時間（例えば、年単位で）空くような場合に、探索パラメータの値を車両の状態に応じて徐々に変化させることができる。

【００５２】

探索パラメータは、例えば、所定の形式のテーブルによって、イベントと関連付けられていてよい。図１１は、イベントと探索パラメータの値との対応付けを説明するための図である。この例では、イベント１～イベント３（イベント１１０１の列）に対して、それぞれの探索パラメータの値（探索パラメータ１１０２の列）が関連付けられている。車両の製造完了時が１つ目のイベントとして定義され、探索パラメータの値がそれ以前より低下するように設定されている。そして、この例では、車両が所定の走行距離の閾値以上走行すると、探索パラメータが段階的に引き下げられ、例えば、最終的にはゼロに設定される。

【００５３】

データ収集サーバ１１０は、当該テーブルを予め記憶しておき、車両１００からイベントの情報を受信すると、当該テーブルを参照して対応する探索パラメータの値を取得し、車両１００に送信する。車両１００は、データ収集サーバ１１０から受信した探索パラメータをモデル処理部２１４に設定して、強化学習アルゴリズムを実行する。

【００５４】
＜車両におけるダンパ制御処理の一連の動作＞
次に、車両におけるダンパ制御処理の一連の動作について、図７を参照して説明する。なお、本処理は、図３の説明において時刻ｔのフィードバックデータが得られた時点から開始される。なお、モデル処理部２１４の動作は、例えば５Ｈｚの動作周波数で行われるものとする。また、本処理では、例えば、初期の探索パラメータがモデル処理部２１４に設定されている。更に、モデル処理部２１４およびアクター４０１などの制御部２００内の構成による処理は、ＣＰＵ２１０がＲＯＭ２１２に格納されたプログラムをＲＡＭ２１１に展開、実行することにより実現される。
【００５５】
Ｓ７０１において、アクター４０１は、データ入力部２１３からフィードバックデータを受け付けて、方策π（ｓ，ａ）に基づき行動（a_t）を出力する。このとき、モデル処理部２１４は、アクター４０１の出力した行動（出力層４１３に相当）から、設定されている探索パラメータに応じて、行動を選択する。そして選択した行動に対応する制御変数を出力する。
【００５６】
Ｓ７０２において、ダンパ制御部１０６は、モデル処理部２１４からの制御変数を受け付けると、ダンパ制御部１０６の内部で用いられている制御変数を、モデル処理部２１４から取得した新たな制御変数に置き換える。そして、ダンパ制御部１０６は、置き換えた制御変数をフィードバックデータに適用することにより、ダンパ１０７の特性を制御する。なお、図７に示すフローチャートでは、簡単のため、Ｓ７０２とＳ７０３のステップは、時刻ｔに対して１回分の制御として記載されている。しかし、ダンパ制御部１０６は、例えば１ＫＨｘの速度で取得可能なフィードバックデータに対し、ダンパ特性を、例えば、１００Ｈｚの動作周波数で制御し、当該動作周波数で制御量（ダンパ１０７の減衰力を制御するための電流量）を制御することができる。この場合、実際には、時刻ｔ＋１までに、Ｓ７０２とＳ７０３の処理が繰り返され得る。Ｓ７０３において、ダンパ制御部１０６は、算出した制御量（例えば電流量）をダンパに供給してダンパ１０７の特性を制御する。
【００５７】
Ｓ７０４において、センサ部１０１は、時刻ｔ＋１までフィードバックデータを（例えば１ＫＨｚの動作周波数で）取得する。
【００５８】
Ｓ７０５において、データ入力部２１３は、フィードバックデータに上述した前処理を適用する。Ｓ７０６において、報酬決定部２１５は、時刻ｔ＋１におけるフィードバックデータに基づいて、上述した報酬（r_{t+1}）を決定し、クリティック４０２に出力する。Ｓ７０７において、クリティック４０２は、アクター４０１の方策を改善するための、上述した方策改善（例えばＴＤ誤差）を算出して、アクター４０１に入力する。
【００５９】
Ｓ７０８において、アクター４０１は、Ｓ７０７における方策改善に基づいて方策π（ｓ，ａ）を更新する。アクター４０１は、例えば、ｐ（s_t，a_t）をｐ（s_t，a_t）＋βδ_tで置き換えるように方策を更新する。Ｓ７０９のおいて、クリティック４０２は、状態価値関数Ｖ（ｓ）を、例えばＶ（ｓ）＋αδ_t（αはステップサイズパラメータ）で置き換えて更新する。クリティック４０２が状態価値関数を更新すると、その後、本処理は終了する。本実施形態では、時刻ｔから時刻ｔ＋１における動作を例に説明したが、図７に示す一連の動作を繰り返して、所定の条件を満たした場合に一連の処理を終了するようにしてもよい。
【００６０】
＜車両における探索パラメータ設定処理の一連の動作＞
次に、車両における探索パラメータ設定処理の一連の動作について、図８を参照して説明する。なお、本処理は、図３の説明において時刻ｔのフィードバックデータが得られた

時点から開始され、図７を参照して説明したダンパ制御処理と独立して並列に実行される。本処理は、ＣＰＵ２１０がＲＯＭ２１２に格納されたプログラムをＲＡＭ２１１に展開、実行することにより実現される。

【００６１】

Ｓ８０１において、データ入力部２１３は、センサ部１０１からのフィードバックデータに基づいて、上述の加工したフィードバックデータを取得する。このフィードバックデータは実環境における学習データとして収集され、必要に応じて記憶部１０８に一時的に記憶される。Ｓ８０２において、制御部２００は、記憶部１０８に一時的に記憶されたフィードバックデータを学習データとして、順次、データ収集サーバ１１０に送信する。

【００６２】

Ｓ８０３において、イベント検知部２１７は、車両１００において所定のイベントが発生したかを判定する。例えば、イベント検知部２１７は、車両１００における所定の機能のアクティベーションされた場合や、ＲＯＭ２１２に製造完了を示す所定のバージョンを示す情報が記憶された場合に、車両の製造完了を検知する。或いは、ユーザ操作に基づいて、製造完了或いは車両登録に関する情報が入力された場合に、車両の製造完了や車両登録を検知してもよい。

【００６３】

また、イベント検知部２１７は、ＲＯＭ２１２或いは記憶部１０８に記憶される走行距離の情報を参照して、当該走行距離が所定の走行距離を超えている場合に、対応するイベントを検知する。このほか、送信した学習データのデータ量をカウントし、所定のデータ量が超えた場合に、対応するイベントを検出してもよい。或いは、所定の時点（例えば車両１００の初期の車両モデルの販売開始時や、車両１００そのものの製造完了時など）からの経過時間が経過している場合に、対応するイベント検知する。制御部２００は、イベント検知部２１７がイベントを検知した場合、処理をＳ８０４に進め、そうでないと判定した場合には、Ｓ８０１に処理を戻す。

【００６４】

Ｓ８０４において、制御部２００は、検知したイベントを示すイベント情報をデータ収集サーバ１１０に送信する。イベント情報は、例えば、イベントに予め割り当てられているイベントの識別子である。

【００６５】

Ｓ８０５において、探索パラメータ設定部２１６は、データ収集サーバ１１０から送信される探索パラメータを取得して、モデル処理部２１４に設定する。取得する探索パラメータは、例えば、図１１を参照して説明した探索パラメータの値（探索パラメータ１１０２）が含まれる。

【００６６】

Ｓ８０６において、モデル処理部２１４は、ニューラルネットワークの演算を実行し、新たな探索パラメータを用いて行動を選択する。そして、選択した行動に対応する制御変数を出力する。このとき、モデル処理部２１４は、ランダムに選択した行動に基づく制御変数が安全上許容できるか否かを判定し、許容できないと判定した場合には当該制御変数をフィルタアウトすることができる。安全上許容できるか否かの判定は、予め実験等で定めた判定条件を用いるようにすればよい。このようにすれば、実環境においてランダムな行動の選択によって突飛な出力が選択される場合であっても安全な制御を担保することができる。

【００６７】

Ｓ８０７において、制御部２００は、車両制御を終了するかを判定し、終了すると判定する場合、その後、本一連の処理を終了する。そうでない場合には、処理をＳ８０１に戻して処理を繰り返す。

【００６８】

このように、本実施形態では、強化学習を用いて制御を行う車両において、車両のライフサイクルにおけるイベントを検知すると、当該イベントに応じて特定される探索パラメ

ータを、強化学習における探索の割合を調整する値として設定する。そして、設定された探索パラメータに従って強化学習を用いた処理を実行する。このとき、第１のイベントが検知された場合、第１のイベントの後である第１の期間に設定される探索の割合を、第１のイベントが検知される前の第２の期間に設定されていた探索の割合より小さくする探索パラメータを設定する。このようにすることで、車両の制御において、車両のライフサイクルにおいて強化学習における探索と活用を継続的に両立することが可能になる。

【００６９】

＜データ収集サーバの構成＞

次に、情報処理サーバの一例としてのデータ収集サーバの機能構成例について、図９を参照して説明する。なお、以降の図を参照して説明する機能ブロックの各々は、統合されまたは分離されてもよく、また説明する機能が別のブロックで実現されてもよい。また、ハードウェアとして説明するものがソフトウェアで実現されてもよく、その逆であってもよい。

【００７０】

制御部９００は、例えば、ＣＰＵ９１０、ＲＡＭ９１１、ＲＯＭ９１２を含み、データ収集サーバ１１０の各部の動作を制御する。制御部９００は、ＣＰＵ９１０がＲＯＭ９１２に格納されたコンピュータプログラムを、ＲＡＭ９１１に展開、実行することにより、制御部９００を構成する各部の機能を発揮させる。

【００７１】

イベント情報取得部９１３は、車両１００から送信されたイベント情報を（通信部９０１を介して）取得する。探索パラメータ制御部９１４は、イベント情報取得部９１３によって取得されたイベントに対応する探索パラメータを特定する。探索パラメータ制御部９１４は、特定した探索パラメータを、イベント情報を送信した車両に送信する。

【００７２】

モデル提供部９１５は、車両１００のモデル処理部２１４に設定される強化学習アルゴリズムの学習モデルをバージョンアップする際に、車両１００にモデル情報を提供する。モデル情報は、当該学習モデルのバージョンやニューラルネットワークの重み付けパラメータなどを含む。モデル提供部９１５は、車両１００から収集された学習データを用いてサーバ上で学習モデルの最適化を行って、学習モデルのバージョンアップを行うことができる。

【００７３】

通信部９０１は、例えば通信用回路等を含む通信デバイスであり、例えばインターネットなどのネットワークを通じて、車両と通信する。通信部９０１は、車両から送信されるフィードバックデータ（学習データ）の情報とを受信し、探索パラメータの情報（或いは学習モデルの情報）を車両１００に送信する。電源部９０２は、データ収集サーバ１１０内の各部に電力を供給する。

【００７４】

記憶部９０３は、ハードディスクや半導体メモリなどの不揮発性メモリである。記憶部９０３は、車両から送信された、上述した学習データの情報を格納するＤＢ１１１を含む。

【００７５】

＜データ収集サーバにおける探索パラメータ送信処理の一連の動作＞

次に、図１０を参照して、データ収集サーバ１１０における探索パラメータ送信処理の一連の動作について説明する。なお、本処理は、制御部９００のＣＰＵ９１０が、ＲＯＭ９１２に記憶されたプログラムをＲＡＭ９１１に展開、実行することにより実現される。

【００７６】

Ｓ１００１において、イベント情報取得部９１３は、車両１００から送信された学習データを通信部９０１を介して取得し、記憶部９０３のＤＢ１１１に蓄積する。Ｓ１００２において、制御部９００は、車両１００からイベント情報を受信したかを判定する。制御部９００は、イベント情報を受信した場合、Ｓ１００３に処理を進め、そうでない場合、Ｓ１００１に処理を戻す。

【００７７】
Ｓ１００３において、探索パラメータ制御部９１４は、イベント情報取得部９１３によって取得されたイベントに対応する探索パラメータを特定する。例えば、予め定められたイベントＩＤに基づいて、イベントに関連付けられた探索パラメータの値を特定する。
【００７８】
Ｓ１００４において、探索パラメータ制御部９１４は、特定した探索パラメータを、イベント情報を送信した車両１００に送信する。データ収集サーバ１１０は、車両１００に探索パラメータを送信すると、その後、本処理を終了する。
【００７９】
このように、データ収集サーバ１１０は、車両から送信されたイベント情報に基づいて強化学習の探索の確率を定義する探索パラメータを特定し、特定した探索パラメータを車両に提供するようにした。このようにすることで、データ収集サーバ１１０は、実環境において走行している多数の車両の探索パラメータの制御を一元管理することが可能になる。
【００８０】
＜その他の実施形態＞
上述の実施形態では、車両１００の制御部２００において、フィードバックデータを取得し、強化学習を用いて方策を算出したうえで探索の確率に応じた方策を選択し、制御変数を出力するようにした。しかしながら、当該制御部２００の処理をデータ収集サーバ１１０側で行ってもよい。すなわち、車両がフィードバックデータをデータ送信サーバに送信する。データ収集サーバ１１０は、受信したフィードバックデータに対し強化学習を用いて方策を算出したうえで探索の確率に応じた方策を選択し、当該方策に応じた制御変数を車両１００に対して出力する。この場合、図７を参照して説明した各ステップ、及び、図８を参照して説明した各ステップを、データ収集サーバ１１０の制御部９００が実施すればよい。Ｓ８０３におけるイベント検出は、イベント検知に必要な情報を車両１００から受信してもよい。例えば、データ収集サーバ１１０がイベントの検知部を備え、車両から製造完了や車両登録の情報を受信したり、車両からの学習データのデータ量をカウントしたり、所定の時点からの経過時間をカウントしてもよい。
【００８１】
上述の実施形態では、車両１００が検出したイベント情報をデータ収集サーバ１１０に送信し、サーバ側でイベント情報に基づく探索パラメータを特定するようにした。しかし、本実施形態は、この例に限定されず、車両１００が、検出したイベントに基づいて、探索パラメータを特定するようにしてもよい。この場合、車両１００は、イベントと探索パラメータとを関連付けたテーブルを、例えば、ＲＯＭ２１２などに記憶しておき、イベントの発生を検知したことに応じて、当該テーブルを参照して探索パラメータを特定してもよい。このようにすれば、車両内において、イベントに応じた探索パラメータの制御が完結する利点がある。
【００８２】
また、上述の実施形態では、データ収集サーバ１１０が、全ての車両に共通である、予め定められたイベントと探索パラメータに係るテーブルを用いて、受信したイベントに対する探索パラメータを特定した。これに対し、データ収集サーバ１１０は、上記イベントと探索パラメータとを関連付けたテーブルを、個別の車両ごとに管理し、個別の車両ごとに、イベントに対する探索パラメータの値が異なるようにしてもよい。一例として、図１１に示した例のように、イベントが走行距離に関するものである場合（例えば、所定の走行距離(TH1)以上走行）、当該イベント発生までに要した時間が所定時間より長い場合には、探索パラメータが少なくなるように補正するようにしてもよい。例えば、標準的な期間でイベント２が発生する車両よりも、より長時間かかってイベント２が発生した場合には探索パラメータの値を０．０２より小さくなるように補正する。こうすることで、いつまでも探索パラメータの値が大きく設定される車両の数を減らすような、個別の車両の状態に応じたコントロールを実現することができる。
【００８３】

或いは、データ収集サーバ１１０は、探索パラメータの値を、車両のモデル（型式）に応じて異ならせてもよい。既に類似する車両のモデルについて十分なデータが収集されており、その車両のモデルに用いる強化学習アルゴリズムの性能が十分に最適化されている場合には、対象モデルの探索パラメータを小さく設定してもよい。

【００８４】

＜実施形態のまとめ＞

１．上記実施形態の制御装置（例えば、２００或いは１００）は、

所定システムのライフサイクルにおけるイベントを検知する検知手段（例えば、２１７）と、

イベントが検知されたことに応じて、検出されたイベントに応じて特定される探索パラメータを、強化学習における探索の割合を調整する値として設定する設定手段（例えば、２１６）と、

設定された探索パラメータに従って強化学習を用いた所定システムに対する所定の制御を実行する処理手段（例えば、２１４）と、を有し、

設定手段は、第１のイベントが検知された場合、第１のイベントの後である第１の期間に設定される探索の割合を、第１のイベントが検知される前の第２の期間に設定されていた探索の割合より小さくする探索パラメータを設定する。

【００８５】

この実施形態によれば、所定システム（例えば車両）に対する制御において、所定システムのライフサイクルにおいて強化学習における探索と活用を継続的に両立することが可能になる。

【００８６】

２．上記実施形態では、

設定手段は、第１の期間に設定される探索の割合を非ゼロとする探索パラメータを設定する。

【００８７】

この実施形態によれば、イベントが検知された後の期間であっても、引き続き学習データの収集において探索的な要素を残すことができる。

【００８８】

３．上記実施形態では、

設定手段は、第１の期間に設定される探索の割合と、第２の期間に設定されていた探索の割合とが非連続となる探索パラメータを設定する。

【００８９】

この実施形態によれば、探索パラメータをイベントの発生に応じて段階的に引き下げることができる。

【００９０】

４．上記実施形態では、

イベントは、所定システムに対する手続きの完了、所定システムの特定の使用状態への到達、及び、所定システムを制御する構成要素の更新の少なくともいずれかを含む。

【００９１】

この実施形態によれば、所定システム（例えば車両）のライフサイクルにおける多様な種類のイベントを扱うことができ、これらのイベントの発生に応じて、探索の割合を変化させることができる。

【００９２】

５．上記実施形態では、

車両に対する手続きの完了は、所定システムの製造完了、及び、所定システムの登録完了の少なくともいずれかを含む。

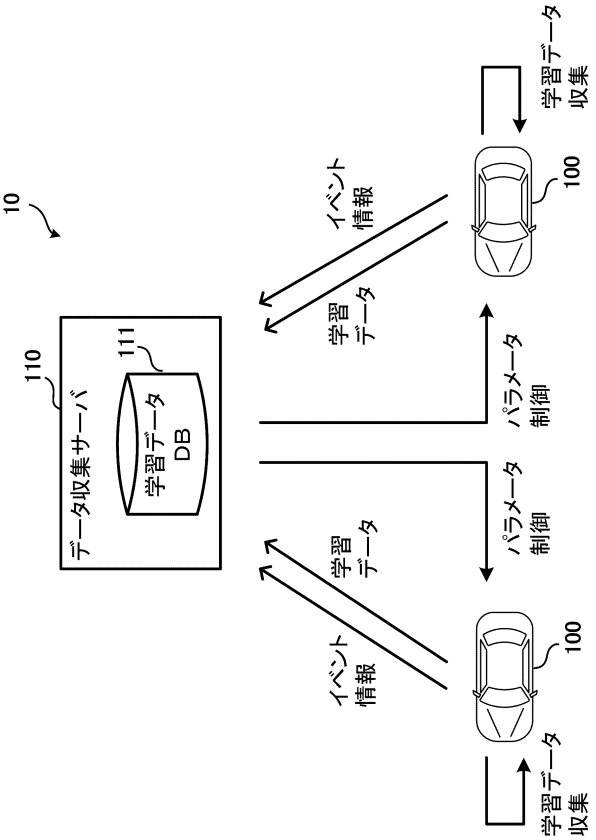
【００９３】

この実施形態によれば、いくつかの所定システム（例えば車両）に対する手続に応じて、探索パラメータを制御することができる。例えば、車両の製造を完了すると、それ以前

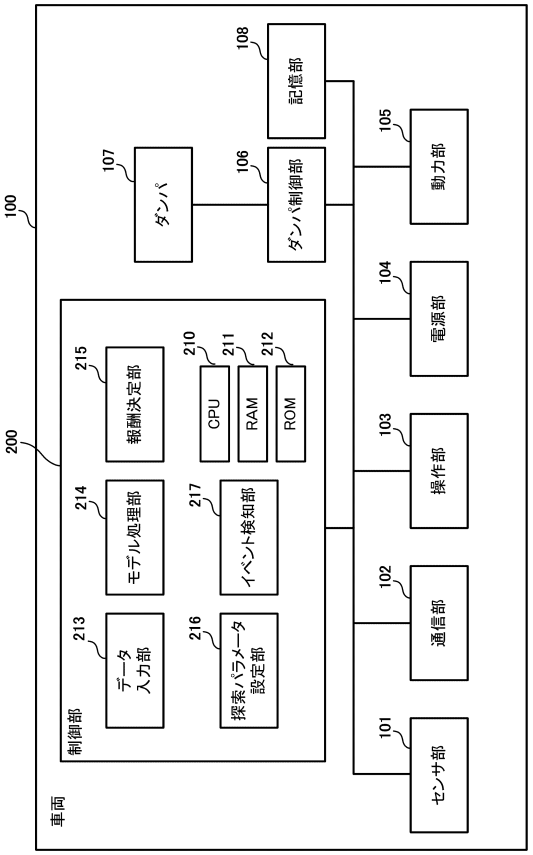
の開発段階で設定していた探索パラメータを、イベント後に低下させることができる。
【００９４】
６．上記実施形態では、
所定システムの特定の使用状態への到達は、所定の時点からの所定日数の経過、所定の時点からの所定走行距離の走行の少なくともいずれかを含む。
【００９５】
この実施形態によれば、経過日数や走行距離などの所定システムの特定の使用状態に応じて、探索パラメータを制御することができる。
【００９６】
７．上記実施形態では、
所定システムを制御する構成要素の更新は、強化学習に用いられる学習モデルのバージョンの更新を含む。
【００９７】
この実施形態によれば、強化学習に用いられる学習モデルのバージョンの更新に応じて、探索パラメータを制御することができる。
【００９８】
８．上記実施形態では、
検出されたイベントに応じて、前記探索パラメータを特定する特定手段を更に有する。
【００９９】
この実施形態によれば、所定システムにおいて、イベントに応じた探索パラメータを特定することができる。
【０１００】
９．上記実施形態では、
検出されたイベントを外部サーバに送信する送信手段（例えば、１０２）と、
イベントに応じて特定された探索パラメータを外部サーバから受信する受信手段（例えば、１０２）と、を更に有する。
【０１０１】
この実施形態によれば、イベントに応じた探索パラメータを外部サーバにおいて行うことができ、車両にある計算機リソースを節約することができる。
【０１０２】
１０．上記実施形態では、
探索パラメータは、所定システムごと、又は所定システムのモデルごとに異なる。
【０１０３】
この実施形態によれば、探索と活用の両立を個別の所定システム（例えば車両）ごと、又は所定システムのモデルごとに変化させることで、個々の所定システムの使用方法や、所定システムのモデルの特性に応じた探索パラメータを設定することができる。
【０１０４】
１１．　上記実施形態では、
処理手段によって実行される強化学習のモデルに対する入力情報と出力情報とを、学習データとして外部サーバに提供する。
この実施形態によれば、外部サーバに、強化学習の学習に有用である利用可能なばらつきのあるデータを送信することができる。
【０１０５】
発明は上記の実施形態に制限されるものではなく、発明の要旨の範囲内で、種々の変形・変更が可能である。
【符号の説明】
【０１０６】
２００…制御部、２１４…モデル処理部、２１６…探索パラメータ設定部、２１７…イベント検知部

【図面】

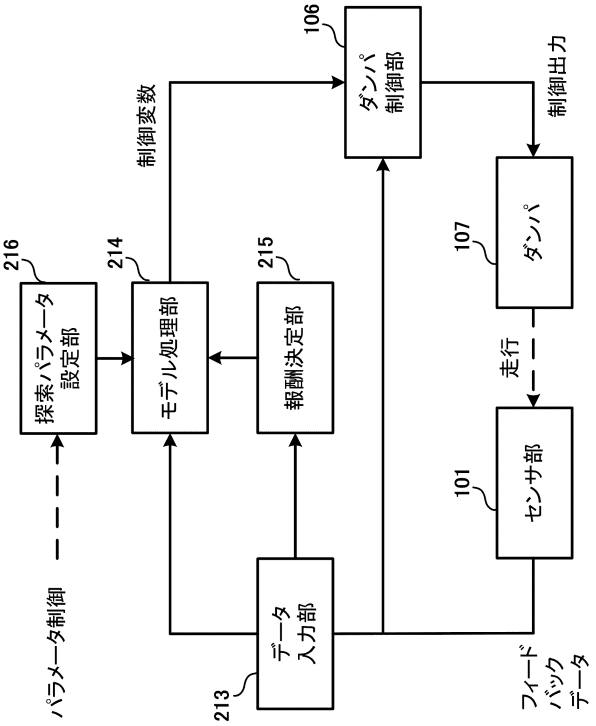
【図１】


10
データ収集サーバ
110
学習データ DB
111
学習データ
収集
イベント情報
学習データ
パラメータ制御
100

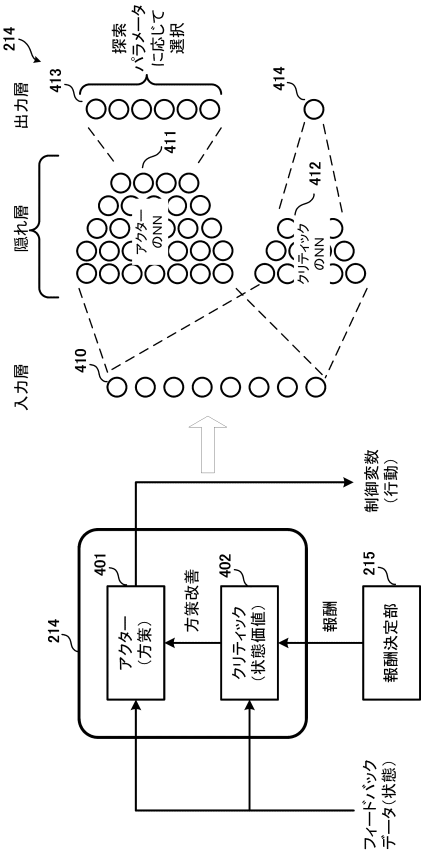
【図２】


車両
100
制御部
200
データ入力部
213
探索パラメータ設定部
216
モデル処理部
214
報酬決定部
215
イベント検知部
217
CPU
210
RAM
211
ROM
212
センサ部
101
通信部
102
操作部
103
電源部
104
動力部
105
ダンパ制御部
106
ダンパ
107
記憶部
108

【図３】


パラメータ制御
探索パラメータ設定部
216
モデル処理部
214
報酬決定部
215
ダンパ制御部
106
制御変数
制御出力
ダンパ
107
走行
センサ部
101
データ入力部
213
フィードバックデータ

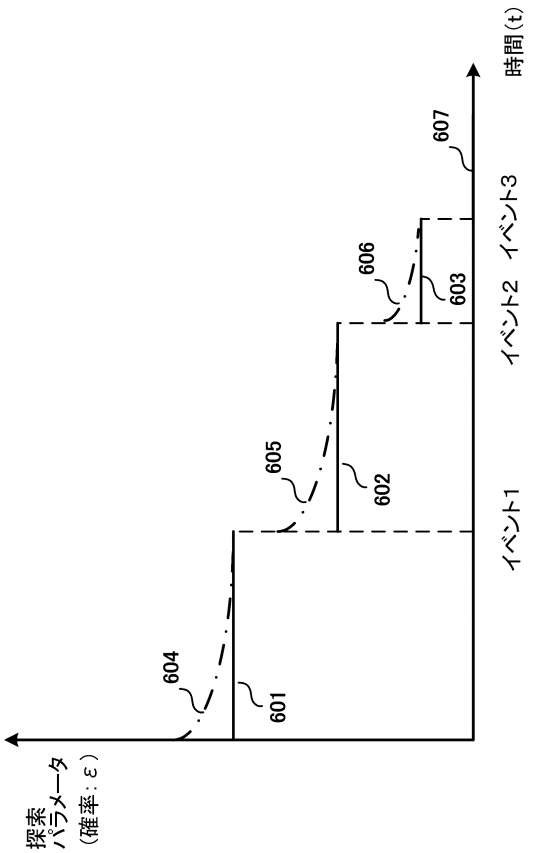
【図４】


214
出力層
413
探索パラメータに応じて選択
411
414
隠れ層
アクターのNN
412
クリティックのNN
入力層
410
制御変数（行動）
401
アクター（方策）
方策改善
402
クリティック（状態価値）
報酬
215
報酬決定部
フィードバックデータ（状態）

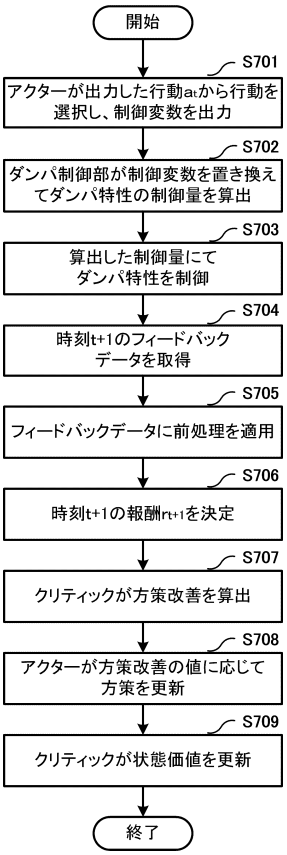
【図５】

計測されるセンサデータ例	センサ
車速	車速センサ
車両ボディの加速度	X, Y, Z方向の加速度センサ
車両ボディの変位	X, Y, Z方向の加速度センサ
車両ボディの角速度	X, Y, Z方向の変位センサ
タイヤ位置（バネ下）の加速度	X, Y, Z方向の加速度センサ
バネの伸縮速度	X, Y, Z方向の速度センサ
バネ伸縮変位	X, Y, Z方向の変位センサ
サスペンション（ダンパ）ストローク速度	サスペンション変位センサ
サスペンション（ダンパ）ストローク変位	サスペンション変位センサ
ステアリング入力	操舵角センサ
加減速	アクセル開度、ブレーキ踏力センサ
サスペンションのコンプライアンス	ロードセル
ＧＰＳデータ（自己位置）	ＧＰＳ
タイヤ空気圧	タイヤ空気圧センサ
各種ダイナミクス制御デバイスの動作	各種ダイナミクス制御デバイスの制御信号
距離画像	LIDAR
前方画像	前方撮影用カメラ

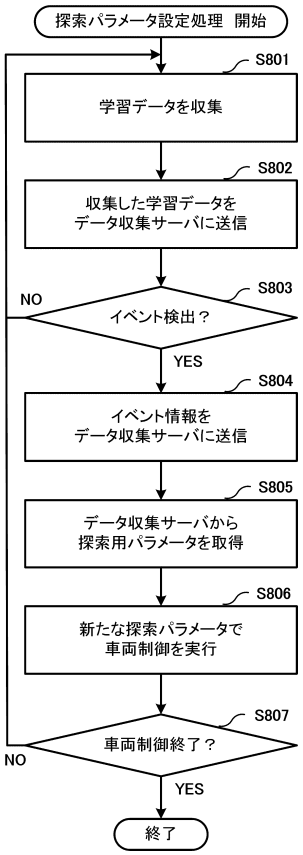
【図６】



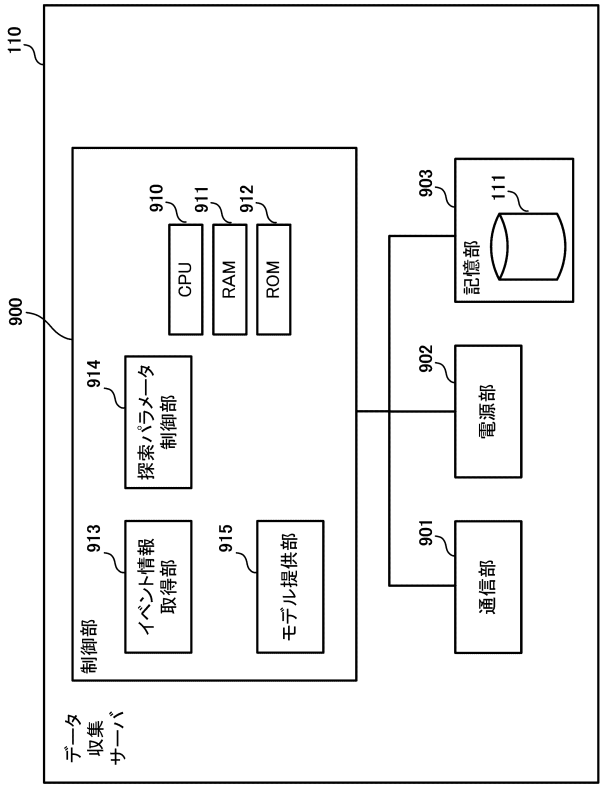
【図７】



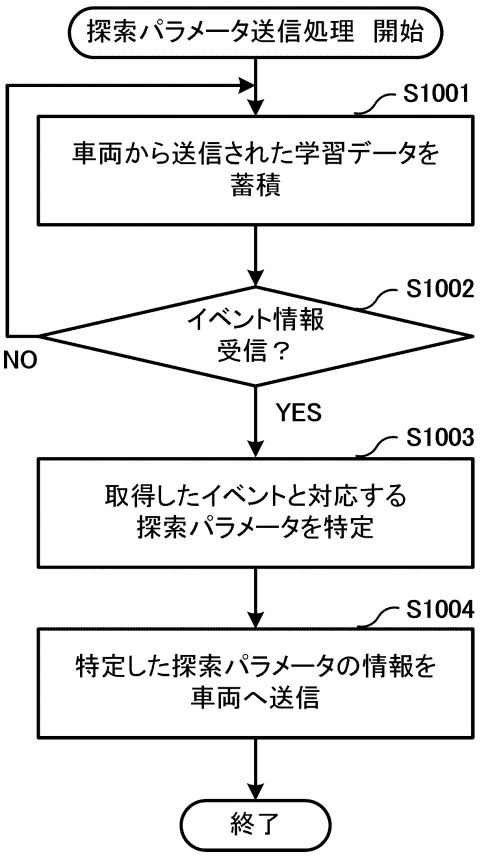
【図８】



【図９】



【図１０】



【図１１】

1101 イベント	説明	1102 探索パラメータ
	（車両製造完了より前の初期値）	0.100
1	車両製造完了時	0.050
2	所定の走行距離（TH1）以上走行	0.020
3	所定の走行距離（TH2）以上走行	0.000

フロントページの続き

(74)代理人　100166648
弁理士　鎗田 伸宜
(72)発明者　藤元 岳洋
埼玉県和光市中央１丁目４番１号 株式会社本田技術研究所内
審査官　大古 健一
(56)参考文献　特開２０１８－１５２０１２（ＪＰ，Ａ）
特開２０１８－１５１８７６（ＪＰ，Ａ）
特開２０１７－１６７８６６（ＪＰ，Ａ）
特開平１０－３２８９８０（ＪＰ，Ａ）
米国特許出願公開第２０１８／０１６５６０２（ＵＳ，Ａ１）
(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０５Ｂ　１／００　－　７／０４
Ｇ０５Ｂ　１１／００　－１３／０４
Ｇ０５Ｂ　１７／００　－１７／０２
Ｇ０５Ｂ　２１／００　－２１／０２
Ｇ０６Ｎ　２０／００
Ｇ０５Ｂ　２３／００　－２３／０２