(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7036545号
(P7036545)

(45)発行日　令和4年3月15日(2022.3.15)

(24)登録日　令和4年3月7日(2022.3.7)

(51)国際特許分類
Ｇ０５Ｂ　13/02　(2006.01)

ＦＩ
Ｇ０５Ｂ　13/02　Ｊ

請求項の数　19　外国語出願　（全36頁）

(21)出願番号　特願2017-131700(P2017-131700)
(22)出願日　平成29年7月5日(2017.7.5)
(65)公開番号　特開2018-37064(P2018-37064A)
(43)公開日　平成30年3月8日(2018.3.8)
審査請求日　令和2年7月3日(2020.7.3)
(31)優先権主張番号　15/205,558
(32)優先日　平成28年7月8日(2016.7.8)
(33)優先権主張国・地域又は機関
米国(US)

(73)特許権者　507342261
トヨタ モーター エンジニアリング アンド マニュファクチャリング ノース アメリカ，インコーポレイティド
アメリカ合衆国、７５０２４ テキサス州、プレイノ、ダブリュ１－３シー・ヘッドクォーターズ・ドライブ、６５６５
(74)代理人　100099759
弁理士　青木 篤
(74)代理人　100123582
弁理士　三橋 真二
(74)代理人　100092624
弁理士　鶴田 準一
(74)代理人　100147555
弁理士　伊藤 公一

最終頁に続く

(54)【発明の名称】　能動的探索なしの強化学習に基づくオンライン学習法及び車両制御方法

(57)【特許請求の範囲】
【請求項１】
車両の自律的動作を適応的に制御するコンピュータ実行型方法であって、該方法は、
ａ）車両を自律的に制御するように構成されたコンピュータ処理システムにおけるcriticネットワークにおいて、受動的に収集されたデータのサンプルと、状態コストとを用いて、推定平均コストと、actorネットワークにより適用されたときに車両の到達コストに対する最小値を生成する近似された到達コスト関数とを決定することと、
ｂ）コンピュータ処理システム内においてcriticネットワークに対して作用的に連結されたactorネットワークにおいて、車両に対して適用されて到達コストに対する最小値を生成する制御入力を決定すること、とを備え、
前記actorネットワークは、推定平均コストと、近似された到達コスト関数から決定された推定到達コストと、車両の現在の状態に対する制御用動力学的値と、受動的に収集されたデータのサンプルとを用いて、ノイズレベルを推定することにより制御入力を決定するように構成され、
前記近似された到達コスト関数は、以下の関係に従い、重み付けされた放射基底関数の線形結合を用いて決定され、
【数１】

$$\hat{Z}(\mathrm{x}) := \sum_{j=0}^{N} \omega_j f_j(\mathrm{x})$$

式中、ωは重みであり、f_jは第ｊ番目の放射基底関数であり、Ｎは近似された到達コスト関数を決定するために使用される放射基底関数の個数であり、且つ、

$\hat{Z}(\mathrm{x})$

は近似された到達コスト関数である、方法。

【請求項２】

前記近似された到達コスト関数において使用される重みωは、以下の関係に従って更新され、

【数２】

$$\omega^{i+1} = \tilde{\omega}^i + \lambda_1 + \sum_{l=0}^{N} \lambda_{2l}\delta_{jl} + \lambda_3 \hat{Z}_{avg} f_k$$

式中、δ_{jj}はディラックのデルタ関数を表し、上付き文字ｉは反復の回数を表し、λ_1、λ_2、λ_3はラグランジュ乗数であり、且つ

$\hat{Z}_{avg}$

は推定平均コストである、請求項１に記載の方法。

【請求項３】

ベルマン方程式の線形化版を用いて決定された近似的な時間差的誤差を用いて前記criticネットワークのパラメータを更新する段階を更に備える、請求項１に記載の方法。

【請求項４】

前記criticネットワークのパラメータを更新する段階は、前記車両が運動しているときに実施される、請求項３に記載の方法。

【請求項５】

車両の自律的動作を適応的に制御するコンピュータ実行型方法であって、該方法は、

ａ）車両を自律的に制御するように構成されたコンピュータ処理システムにおけるcriticネットワークにおいて、受動的に収集されたデータのサンプルと、状態コストとを用いて、推定平均コストと、actorネットワークにより適用されたときに車両の到達コストに対する最小値を生成する近似された到達コスト関数とを決定することと、

ｂ）コンピュータ処理システム内においてcriticネットワークに対して作用的に連結されたactorネットワークにおいて、車両に対して適用されて到達コストに対する最小値を生成する制御入力を決定すること、とを備え、

前記actorネットワークは、推定平均コストと、近似された到達コスト関数から決定された推定到達コストと、車両の現在の状態に対する制御用動力学的値と、受動的に収集されたデータのサンプルとを用いて、ノイズレベルを推定することにより制御入力を決定するように構成され、

当該方法は、ベルマン方程式の線形化版を用いて決定された近似的な時間差的誤差を用い

て前記criticネットワークのパラメータを更新する段階を更に備え、
前記criticネットワークにより決定された推定平均コストは、以下の関係に従って更新され、
【数３】

$$\hat{Z}_{avg}^{i+1} = \hat{Z}_{avg}^{i} - \alpha_2{}^{i} e_k \hat{Z}_k$$

式中、βは学習率であり、e_kは近似的な時間差的誤差であり、

$\hat{Z}_k$

は前記近似された到達コスト関数から決定された推定コストであり、

$\hat{Z}_{avg}^{i}$

は状態ｉにおける推定平均コストであり、且つ、

$\hat{Z}_{avg}^{i+1}$

は状態ｉ＋１における推定平均コストである、方法。
【請求項６】
前記受動的に収集されたデータは、前記criticネットワークのパラメータを更新する間に使用される唯一のデータである、請求項３に記載の方法。
【請求項７】
車両の自律的動作を適応的に制御するコンピュータ実行型方法であって、該方法は、
ａ）車両を自律的に制御するように構成されたコンピュータ処理システムにおけるcriticネットワークにおいて、受動的に収集されたデータのサンプルと、状態コストとを用いて、推定平均コストと、actorネットワークにより適用されたときに車両の到達コストに対する最小値を生成する近似された到達コスト関数とを決定することと、
ｂ）コンピュータ処理システム内においてcriticネットワークに対して作用的に連結されたactorネットワークにおいて、車両に対して適用されて到達コストに対する最小値を生成する制御入力を決定すること、とを備え、
前記actorネットワークは、推定平均コストと、近似された到達コスト関数から決定された推定到達コストと、車両の現在の状態に対する制御用動力学的値と、受動的に収集されたデータのサンプルとを用いて、ノイズレベルを推定することにより制御入力を決定するように構成され、
当該方法は、以下の関係に従って決定された近似的な時間差的誤差を用いて前記criticネットワークのパラメータを更新する段階を更に備え、
【数４】

$$e_k := \hat{Z}_{avg}\hat{Z}_k - \exp(-q_k)\hat{Z}_{k+1}$$

式中、e_kは近似的な時間差的誤差であり、

$\hat{Z}_{avg}$
は推定平均コストであり、

$\hat{Z}_k$

は状態ｋにおける推定到達コストであり、

$\hat{Z}_{k+1}$

は状態ｋ＋１における推定到達コストであり、且つ、q_kは状態ｋにおける状態コストである、方法。
【請求項８】
前記近似された到達コスト関数は、前記criticネットワークにおいてリアルタイムで学習される、請求項１に記載の方法。
【請求項９】
車両の自律的動作を適応的に制御するコンピュータ実行型方法であって、該方法は、
ａ）車両を自律的に制御するように構成されたコンピュータ処理システムにおけるcriticネットワークにおいて、受動的に収集されたデータのサンプルと、状態コストとを用いて、推定平均コストと、actorネットワークにより適用されたときに車両の到達コストに対する最小値を生成する近似された到達コスト関数とを決定することと、
ｂ）コンピュータ処理システム内においてcriticネットワークに対して作用的に連結されたactorネットワークにおいて、車両に対して適用されて到達コストに対する最小値を生成する制御入力を決定すること、とを備え、
前記actorネットワークは、推定平均コストと、近似された到達コスト関数から決定された推定到達コストと、車両の現在の状態に対する制御用動力学的値と、受動的に収集されたデータのサンプルとを用いて、ノイズレベルを推定することにより制御入力を決定するように構成され、
前記ノイズレベルは、以下の関係に従い、重み付けされた基底関数の線形結合を用いて学習され、
【数５】

$$\rho(\mathrm{x};\mu) \approx \hat{\rho}(\mathrm{x};\mu) := \sum_{j}^{M} \mu_j g_j(\mathrm{x})$$

式中、ρは推定ノイズレベルであり、μ_jは、g_jにより表された第ｊ番目の放射基底関数に対する重みであり、且つ、Ｍは、ノイズレベルを推定するために使用されるべき放射基底関数の個数である、方法。
【請求項１０】
以下の関係に従って決定された近似誤差を用いて前記actorネットワークの重み付けパラメータを更新する段階を更に備え、
【数６】

$$d_k \approx q_k \Delta t - \hat{V}_{k+1} + \hat{V}_k + \hat{V}_{avg} + L_{k,k+1}\rho_k,$$

式中、ｄｋは近似誤差であり、ｑｋは状態ｋにおける状態コストであり、

$\hat{V}_k$

は状態ｋにおいて近似された到達コストであり、

$\hat{V}_{k+1}$

は状態ｋ＋１において近似された到達コストであり、

$\hat{V}_{avg}$

は近似された平均コストであり、且つ、
【数７】

$$L_{k,k+1} := \left(0.5\hat{V}_{\mathrm{k}} - \hat{V}_{\mathrm{k+1}}\right)^{\mathrm{T}} B_k B_k^{\mathrm{T}} \hat{V}_k \Delta t$$

であり、式中、Ｂｋは状態ｋにおける制御用動力学的値である、請求項９に記載の方法。
【請求項１１】
前記actorネットワークの重み付けパラメータの更新は、前記車両が運動しているときに実施される、請求項１０に記載の方法。
【請求項１２】
前記actorネットワークの重み付けパラメータは、以下の関係に従って更新され、
【数８】

$$\mu^{i+1} = \mu^i - \beta^i d_k L_{k,k+1} g_k,$$

式中、

μ^{i+1}

は状態ｉ＋１における重み付けパラメータの値であり、

μ^i

は状態ｉにおける重み付けパラメータの値であり、βｉは学習率であり、ｄｋは時間差的誤差であり、且つｇは放射基底関数である、請求項１０に記載の方法。
【請求項１３】

前記制御入力を用い、前記自律的動作を制御すべく使用可能な制御ポリシーを修正する段階を更に備える、請求項１に記載の方法。
【請求項１４】
前記推定平均コストが収束するまで、前記段階（ａ）及び（ｂ）を反復的に実施して前記制御入力を再決定することにより、前記自律的動作を制御するために使用可能な制御ポリシーを最適化する段階を更に備える、請求項１に記載の方法。
【請求項１５】
前記制御ポリシーは、能動的探索なしで最適化される、請求項１４に記載の方法。
【請求項１６】
車両の自律的動作を適応的に制御するように構成されたコンピュータ処理システムであって、該コンピュータ処理システムは、該コンピュータ処理システムの動作を制御する一つ以上のプロセッサと、該一つ以上のプロセッサにより使用可能なデータ及びプログラム命令を記憶するメモリとを備え、
前記一つ以上のプロセッサは、前記メモリ内に記憶された命令を実行して、
ａ）受動的に収集されたデータのサンプルと、状態コストとを用いて、推定平均コストと、前記車両の到達コストに対する最小値を生成する近似された到達コスト関数とを決定し、且つ、
ｂ）前記車両に対して適用されて前記到達コストに対する最小値を生成する制御入力を決定する、ように構成され、
前記一つ以上のプロセッサは、前記推定平均コストと、前記近似された到達コスト関数から決定された到達コストと、前記車両の現在の状態に対する制御用動力学的値と、受動的に収集されたデータのサンプルとを用いて、ノイズレベルを推定することにより制御入力を決定するように構成され、
前記近似された到達コスト関数は、以下の関係に従い、重み付けされた放射基底関数の線形結合を用いて決定され、
【数１】

$$\hat{Z}(x) := \sum_{j=0}^{N} \omega_j f_j(x)$$

式中、ωは重みであり、f_jは第ｊ番目の放射基底関数であり、Ｎは近似された到達コスト関数を決定するために使用される放射基底関数の個数であり、且つ、

$\hat{Z}(x)$

は近似された到達コスト関数である、コンピュータ処理システム。
【請求項１７】
前記一つ以上のプロセッサは、前記メモリ内に記憶された命令を実行し、前記推定平均コストが収束するまで、前記段階（ａ）及び（ｂ）を反復的に実施して前記制御入力を再決定することにより、前記自律的動作を制御するために使用可能な制御ポリシーを最適化するように構成される、請求項１６に記載のコンピュータ処理システム。
【請求項１８】
コンピュータシステムにより実行可能な命令が自身内に記憶された、一時的でないコンピュータ可読媒体であって、
ａ）受動的に収集されたデータのサンプルと、状態コストとを用いて、推定平均コストと、車両の到達コストに対する最小値を生成する近似された到達コスト関数とを決定することと、

ｂ）前記車両に対して適用されて前記到達コストに対する最小値を生成する制御入力を決定すること、とを備える機能を実施させ、
前記制御入力は、前記到達コストに対する最小値を生成し、且つ前記推定平均コストと、前記近似された到達コスト関数から決定された到達コストと、前記車両の現在の状態に対する制御用動力学的値と、受動的に収集されたデータのサンプルとを用いて、ノイズレベルを推定することにより制御入力が決定され、
前記近似された到達コスト関数は、以下の関係に従い、重み付けされた放射基底関数の線形結合を用いて決定され、
【数１】

$$\hat{Z}(x) := \sum_{j=0}^{N} \omega_j f_j(x)$$

式中、ωは重みであり、f_jは第ｊ番目の放射基底関数であり、Ｎは近似された到達コスト関数を決定するために使用される放射基底関数の個数であり、且つ、

$\hat{Z}(x)$

は近似された到達コスト関数である、一時的でないコンピュータ可読媒体。
【請求項１９】
前記命令は、前記推定平均コストが収束するまで、前記段階（ａ）及び（ｂ）を反復的に繰り返して前記制御入力を再決定することにより、自律的動作を制御するために使用可能な制御ポリシーを最適化するように実行可能である、請求項１８に記載の一時的でないコンピュータ可読媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、車両を自律的に制御する方法に関し、更に詳細には、車両の操作を自律的に制御するために使用可能な制御ポリシー（control policy）を修正及び／又は最適化するための強化学習方法に関する。
【背景技術】
【０００２】
一定形式のシステムにおいては、周囲環境を能動的に探索することにより、最適なシステム制御ポリシーを決定するために、モデルフリー（model-free）強化学習（ＲＬ）技術が採用され得る。しかし、車両が採用し得る全ての動作の膨大な能動的探索（active exploration）に伴う潜在的に否定的な結果により、車両の自律的制御に対して使用可能な制御ポリシーに対して従来のＲＬ手法を適用することは困難であり得る。これに加え、車両安全性の確保を支援するために必要とされる態様で能動的探索を行うと、大きなコンピュータ処理コストが必要とされ得る。代替策として、車両が動作している周囲環境の正確な動力学的モデルを利用することにより、能動的探索なしで最適な制御ポリシーを決定すべく、モデルベースのＲＬ技術が採用され得る。しかし、自律車両が動作している複雑な周囲環境は、正確にモデル化することが非常に困難なことがある。
【発明の概要】
【０００３】
本明細書中に記述された実施形態の一つの見地においては、車両の自律的動作を適応的に（adaptively）制御するコンピュータ実行型方法が提供される。該方法は、（ａ）車両を自律的に制御すべく構成されたコンピュータ処理システムにおけるcriticネットワークに

おいて、受動的に収集されたデータのサンプルと、状態コストとを用いて、推定平均コストと、actorネットワークにより適用されたときに車両の到達コストに対する最小値を生成する近似された到達コスト関数とを決定する段階と、（ｂ）コンピュータ処理システム内においてcriticネットワークに対して作用的に連結されたactorネットワークにおいて、車両に対して適用されて到達コストに対する最小値を生成する制御入力を決定する段階とを含み、actorネットワークは、平均コストと、近似された到達コスト関数から決定された到達コストと、車両の現在の状態に対する制御用動力学的値と、受動的に収集されたデータとを用いて、ノイズレベルを推定することにより制御入力を決定すべく構成される。

【０００４】

本明細書中に記述された実施形態の別の見地においては、車両の自律的動作を適応的に制御するように構成されたコンピュータ処理システムが提供される。該コンピュータ処理システムは、該コンピュータ処理システムの動作を制御する一つ以上のプロセッサと、該一つ以上のプロセッサにより使用可能なデータ及びプログラム命令を記憶するメモリと含み、上記一つ以上のプロセッサは、メモリ内に記憶された命令を実行して、（ａ）受動的に収集されたデータのサンプルと、状態コストとを用いて、推定平均コストと、車両の到達コストに対する最小値を生成する近似された到達コスト関数とを決定し、且つ（ｂ）車両に対して適用されて到達コストに対する最小値を生成する制御入力を決定する、ように構成され、一つ以上のプロセッサは、平均コストと、到達コスト関数から決定された到達コストと、車両の現在の状態に対する制御用動力学的値と、受動的に収集されたデータのサンプルとを用いて、ノイズレベルを推定することにより制御入力を決定するように構成される。

【０００５】

本明細書中に記述された実施形態の別の見地においては、一時的でないコンピュータ可読媒体が提供される。該媒体は、コンピュータシステムにより実行可能な命令を該媒体内に記憶し、該コンピュータシステムに、受動的に収集されたデータのサンプルと、状態コストとを用いて、推定平均コストと、車両の到達コストに対する最小値を生成する近似された到達コスト関数とを決定することと、車両に対して適用されて到達コストに対する最小値を生成する制御入力を決定することとを備える機能を実施させ、制御入力は、到達コストに対する最小値を生成し、且つ、平均コストと、到達コスト関数から決定された到達コストと、車両の現在の状態に対する制御用動力学的値と、受動的に収集されたデータのサンプルとを用いて、ノイズレベルを推定することにより制御入力が決定される。

【図面の簡単な説明】

【０００６】

【図１】本明細書中に記述された実施形態に係る、（例えば自律車両などの）システムに対する制御入力を決定すべく且つシステム制御ポリシーを修正及び／又は最適化すべく構成されたコンピュータ処理システムのブロック図である。

【図２】本明細書中に記述された方法に係る、車両制御入力の決定、及び／又は、制御ポリシーの修正若しくは最適化の間における情報の流れを示す概略図である。

【図３】制御入力を決定し且つ制御ポリシーを修正及び／又は最適化する方法の実施形態の動作を示すフローチャートである。

【図４】本明細書中に記述された実施形態に係る、一つ以上の制御入力と制御ポリシーとを使用する自律的制御に向けて構成された車両であって、当該車両に対する制御入力を決定すべく且つ自律車両操作制御ポリシーを修正及び／又は最適化すべく構成されたコンピュータ処理システムが組み込まれた車両の概略的ブロック図である。

【図５】本明細書中に記述された実施形態に係る方法を用いる、高速道路合流用の制御ポリシーの最適化の例において採用された車両の構成の概略図である。

【図６】図５に示された車両の構成に関して実施される最適化のグラフ表示である。

【発明を実施するための形態】

【０００７】

本明細書中に記述された実施形態は、コンピュータ実行型の強化学習（ＲＬ）方法に関す

るものであり、この強化学習方法は、車両を自律的に制御すべく使用可能な制御入力を決定するため、及び車両の操作を自律的に制御する制御ポリシーを修正及び／又は最適化すべく制御入力を使用するために使用可能である。この方法は、（例えば、動作を実施し、且つその動作の結果を監視して制御ポリシーを決定及び改変することを伴い得る）能動的探索を使用せずに、制御入力を決定し且つ制御ポリシーを最適化し得る。本明細書中に記述された方法は、能動的探索の代わりに、受動的に収集されたデータと、部分的に既知であるシステムの動力学的モデルと、制御されている車両に関する既知の制御用動力学的モデルとを使用する。

【０００８】

本開示に関連して、「オンライン」とは、コンピュータ処理システムが学習し得ると共に、actor及びcriticのネットワークパラメータが、上記システムが動作するにつれて（例えば車両が移動するなどにつれて）、コンピュータ処理され且つ更新され得ることを意味する。オンラインのソルーションを用いてactorパラメータ及びcriticパラメータを決定かつ更新すると、車両及びシステムの動力学的値（dynamics）の変更が許容され得る。同様に、自律的動作とは、自律的に実施される動作である。

【０００９】

図１は、本明細書中に開示される種々の実施形態に係る方法を実現すべく構成されたコンピュータ処理システム１４のブロック図である。更に詳細には、少なくとも一つの実施形態において、コンピュータ処理システム１４は、本明細書中に記述された方法に従い、制御入力を決定すべく構成され得る。また、コンピュータ処理システムは、システム（例えば、自律車両）を制御して特定の操作若しくは機能を自律的に実施すべく使用可能な制御ポリシーを修正及び／又は最適化するようにも構成され得る。

【００１０】

少なくとも一つの実施形態において、コンピュータ処理システムは、車両に組み込まれ得ると共に、生成された制御入力を使用して車両の操作の制御に向けられた制御ポリシーを修正及び／又は最適化すべく構成され得る。制御入力を決定するため及び制御ポリシーを修正及び／又は最適化するためにコンピュータ処理システムにより必要とされる情報（例えば、データ、命令、及び／又は他の情報）の少なくとも幾つかは、任意の適切な手段から、例えば車両センサから又は無線接続を介して遠隔データベースのような車外情報源から、受信され且つ／又はそれにより収集され得る。幾つかの実施形態においては、制御ポリシーを修正及び／又は最適化するためにコンピュータ処理システムにより必要とされる情報（例えば、データ）の少なくとも幾つかは、車両の操作の前に（例えば、メモリ内に記憶されたデータ及び他の情報として）コンピュータ処理システムに提供され得る。また、コンピュータ処理システムは、制御入力に従って且つ／又は修正若しくは最適化された制御ポリシーに従って車両を制御することで、関連する自律的動作を実施するようにも構成され得る。

【００１１】

少なくとも一つの実施形態において、コンピュータ処理システムは、（例えばスタンドアロンのコンピュータ処理システムとして）車両から遠隔的に配置され得ると共に、制御入力を決定すべく且つ車両の自律的動作の実施に向けられた制御ポリシーを修正及び／又は最適化するように構成され得る。遠隔的なコンピュータ処理システムによって生成された最適化又は修正された制御ポリシーは、車両による展開のために車両のコンピュータ処理システムへロード又はインストールされて、実際の交通環境において車両を制御し得る。

【００１２】

図１を参照すると、コンピュータ処理システム１４は、コンピュータ処理システム１４及び関連する構成要素の全体的な操作を制御する（少なくとも一つのマイクロプロセッサを含み得る）一つ以上のプロセッサ５８であって、メモリ５４のような一時的でない（non-transitory）コンピュータ可読媒体内に記憶された命令を実行する、プロセッサ５８を含み得る。本開示に関連して、コンピュータ可読記憶媒体とは、命令を実行するシステム、装置若しくはデバイスによって使用されるか又はそれに関連して使用されるプログラムを

含む又は記憶し得る任意の有形媒体であり得る。プロセッサ５８は、プログラムコード中に含まれた命令を実施すべく構成された少なくとも一つのハードウェア回路（例えば、集積回路）を含み得る。複数のプロセッサ５８が在る構成において、斯かるプロセッサは相互から独立して作動し得るか、又は、一つ以上のプロセッサが相互に協働して作動し得る。

【００１３】

幾つかの実施形態において、コンピュータ処理システム１４は、ＲＡＭ５０、ＲＯＭ５２、及び／又は他の任意で適切な形態のコンピュータ可読メモリを含み得る。メモリ５４は、一つ以上のコンピュータ可読メモリを備え得る。一つ又は複数のメモリ５４は、コンピュータ処理システム１４の構成要素であり得るか、又は、一つ又は複数のメモリは、コンピュータ処理システム１４に作用的に接続されてコンピュータ処理システム１４に使用され得る。本説明を通して使用される「作用的に接続された」という語句は、直接的な物理接触のない接続を含め、直接的又は間接的な接続を含み得る。

【００１４】

一つ以上の構成において、本明細書中に記述されたコンピュータ処理システム１４は、人工的又はコンピュータ的な知能要素、例えば、ニューラルネットワーク、ファジィ論理回路、又は他の機械学習アルゴリズム、を組み込み得る。更に、一つ以上の構成において、本明細書中に記述された特定の機能又は操作を実施するように構成されたハードウェア及び／又はソフトウェア要素は、複数の要素及び／又は箇所に分散され得る。コンピュータ処理システム１４に加え、車両は、コンピュータ処理システム１４により実施される制御機能を増強若しくは支援するために、又は他の目的のために、付加的なコンピュータ処理システム及び／又はデバイス（図示せず）を組み込み得る。

【００１５】

メモリ５４は、データ６０、及び／又は、プロセッサ５８によって実行されて種々の機能を実行し得る命令（例えば、プログラムロジック）５６を含み得る。データ６０は、受動的に収集されたデータを含み得る。受動的に収集されたデータは、能動的探索から収集されたのではないデータとして定義され得る。受動的に収集されたデータの一例は、ビルの頂部に取付けられたカメラを用いて高速道路の入口の周りにおける車両の軌跡の獲得を記述する、http://www.fhwa.dot.gov/publications/research/operations/06137/に記述されたデータセットである。別の例において、受動的に収集されたデータは、人間の運転者により実行された操縦に応じて車両センサにより収集されたデータを含み得る。人間の運転者により実行された操縦と、この操縦が実行された車両環境条件と、この操縦に引き続き且つ／又はこの操縦に応じて車両周囲において生じた事象と、に関して、データが収集されてコンピュータ処理システムに提供され得る。或いは、コンピュータ処理システムが車両に設置されたときに、コンピュータ処理システム１４は、（制御ポリシー１０１のような）一つ以上の車両制御ポリシーのオンラインでの修正及び／又は最適化のために、斯かる受動的に収集されたデータを収集及び／又は受信するように構成され得る。

【００１６】

車両制御用動力学的モデル８７は、種々の入力に対して車両が如何に応答するかを記述する刺激応答モデル（stimulus-response model）であり得る。車両制御用動力学的モデル８７は、本明細書中に記述されるように、車両に対する制御入力を決定する上で且つ制御ポリシー１０１を修正及び／又は最適化する上で使用されるべく、（受動的に収集されたデータを用いて）所定の車両状態ｘにおける車両に対して状態コストｑ（ｘ）及び制御用動力学的値Ｂ（ｘ）を決定するように使用され得る。与えられた任意の車両に対する車両制御用動力学的モデル８７が決定されて、メモリ内に記憶され得る。

【００１７】

再び図１を参照すると、コンピュータ処理システムの実施形態は、２つの学習システム又は学習ネットワーク、並びに相互に作用するactorネットワーク（又は「actor」）８３及びcriticネットワーク（又は「critic」）８１も含み得る。これらネットワークは、例えば、人工ニューラルネットワーク（ＡＮＮ）を用いて実現され得る。

【００１８】

本明細書中に記述された目的に対し、（変数πによっても表される）制御ポリシー１０１は、一群の車両の状態のうちの各状態ｘに応じて車両により取られるべき動作ｕを特定又は決定する関数又は他の関係として定義され得る。故に、自律的動作の実行中の車両の各状態ｘに対し、車両は、関連する動作ｕ＝π（ｘ）を実施するように制御され得る。したがって、制御ポリシーは、車両の操作を制御して、例えば、高速道路合流などの関連する操作を自律的に実施する。actor８３は、制御ポリシーに関して動作し、criticから受信した情報及び他の情報を用いて、ポリシーを修正及び／又は最適化し得る。制御ポリシーにより自律的に制御された車両操作は、高速道路への合流、又は、車線の変更のような特定の目的を達成すべく実施される一つの運転操作又は一群の運転操作として定義され得る。

【００１９】

コンピュータ処理システム１４は、制御ポリシーの修正及び最適化に対して使用可能である新規な半モデルフリーＲＬ方法（semi-model-free RL method）（本明細書においては受動的actor／critic（ｐＡＣ）方法という）を実行するように構成される。この方法において、criticは、車両の種々の状態に対する評価関数を学習し、且つ、actorは、能動的探索なしで、代わりに受動的に収集されたデータと既知の車両制御用動力学的モデルとを用いて制御ポリシーを改善する。この方法は、部分的に既知であるシステムの動力学的モデルを使用することにより、能動的探索に対する必要性を回避する。この方法は、車両環境の制御されていない動力学的値又は過渡的なノイズレベルに関する知見を必要としない。この方法は、例えば、環境がノイズ的に如何に展開するかのサンプルは入手可能であるが車両センサにより能動的に探索することは困難であり得る自律車両に関して、実行可能である。

【００２０】

制御入力を決定し且つ制御ポリシーを修正及び／又は最適化する目的に対し、状態ｘ∈Ｒⁿ及び制御入力ｕ∈Ｒᵐにより、離散時間確率論的動力学系は以下のように定義され得る。

【数１】

$$\Delta \mathbf{x} = A(\mathbf{x}_t)\Delta t + B(\mathbf{x}_t)\mathbf{u}_t\Delta t + C(\mathbf{x}_t)d\omega \quad (1)$$

式中、ω（ｔ）はブラウニアン運動であり、

$A(\mathbf{x}_t)$

、

$B(\mathbf{x}_t)\mathbf{u}_t$

及び

$C(\mathbf{x}_t)$

は、それぞれ、受動的動力学的値、車両制御用動力学的値、及び、過渡的ノイズレベルである。Δｔは、時間のステップサイズである。この種の系は、多くの状況において生ずる（例えば、ほとんどの機械系のモデルはこれらの動力学に従う）。関数Ａ（ｘ）、Ｂ（ｘ）及びＣ（ｘ）は、理解されるべく、モデル化されている特定の系に依存する。受動的動力学的値は、車両の環境における変化であって、車両システムに対する制御入力の結果ではない変化を含む。

【００２１】

本明細書中に記述された方法及びシステムにおいて、離散時間動力学系に対するＭＤＰは、タプル

$\langle \mathcal{X}, \mathcal{U}, P, R \rangle$

であり、式中、

$\mathcal{X} \subseteq \mathbf{R}^n$

及び

$\mathcal{U} \subseteq \mathbf{R}^m$

は、状態空間及び動作空間である。

$P := \{p(\mathbf{y}|\mathbf{x},\mathbf{u})|\mathbf{x},\mathbf{y} \in \mathcal{X}, \mathbf{u} \in \mathcal{U}\}$

は、動作による状態遷移モデルであり、且つ、

$R := \{r(\mathbf{x},\mathbf{u})|\mathbf{x} \in \mathcal{X}, \mathbf{u} \in \mathcal{U}\}$

は、状態ｘ及び動作ｕに関する即時コスト関数である。先に記述されたように、制御ポリシー

$\mathbf{u} = \pi(\mathbf{x})$

は、状態ｘから動作

$\mathbf{u}$

へとマッピングする関数である。予期される累積コストである、ポリシーπの下での到達コスト関数（cost-to-go function）（又は価値関数）

$V^{\pi}(\mathbf{x})$

は、無限時間区間（infinite horizon）の平均コストの最適性判断基準の下で、以下のように定義される。

【数２】

$$V^{\pi}(x_t) := \sum_{k=1}^{\infty} p\big(\mathbf{x}_k, \pi(\mathbf{x}_k)\big) r\big(\mathbf{x}_k, \pi(\mathbf{x}_k)\big) \Delta t - V_{avg}^{\pi}$$

$$V_{avg}^{\pi} := \lim_{N \to \infty} \frac{1}{N\Delta t} \sum_{k=1}^{N} p\big(\mathbf{x}_k, \pi(\mathbf{x}_k)\big) r(\mathbf{x}_k, \pi(\mathbf{x}_k)) \Delta t$$

式中、

V_{avg}^{π}

は平均コストであり、ｋは時間インデックスであり、且つ、Δｔは時間ステップである。最適な到達コスト関数は、以下の離散時間ハミルトン－ヤコビ－ベルマン方程式を満足する。

【数３】

$$V_{avg}^{\pi} + V^{\pi}(\mathbf{x}_k) = \min_{u_k} Q^{\pi}(\mathbf{x}_k, \mathbf{u}_k \qquad (2)$$

式中、 $Q^{\pi}(\mathbf{x}_k, \mathbf{u}_k) := r(\mathbf{x}_k, \mathbf{u}_k)\Delta t + \mathcal{G}[V^{\pi}](\mathbf{x}_k)$,であり、且つ、

$\mathcal{G}[V^{\pi}](\mathbf{x}_k) := \int_X p(\mathbf{y}|\mathbf{x}_k, \boldsymbol{\pi})\, V^{\pi}(\mathbf{y})d\mathbf{y}$ である。

式中、

$Q^{\pi}(\mathbf{x}, \mathbf{u})$

は動作価値関数であり、且つ、

$\mathcal{G}[\cdot]$

は積分演算子である。ＭＤＰの目的は、以下の関係に従い、無限時間区間に亘り、平均コストを最小化する制御ポリシーを見出すことである。

【数４】

$$\pi^*(\mathrm{x}_k) = \arg\min_{\pi} \mathbb{E}\left[V_{avg}^{\pi}\right]$$

ここで、最適な制御ポリシーにおける値は、上付き文字*を以て表され得る（例えば、

V^*

、

V_{avg}^*

）。

【００２２】

離散時間動力学系に対する線形ＭＤＰ（Ｌ－ＭＤＰ）は、連続的な状態空間及び動作空間に対して厳密な解が迅速に求められ得るという利点を備えた汎用マルコフ決定過程のサブクラスである。構築された動力学的値、及び、別体的な状態コスト及び制御コストの下で、ベルマン方程式は、組み合わされた状態コスト及び制御されていない動力学的値の線形

固有関数を見出すことに解が限定された線形微分方程式として再構築され得る。その後、Ｌ－ＭＤＰに対する到達コスト関数（又は、価値関数）は、正確な動力学的モデルが利用可能であるときに、二次プログラミング（ＱＰ）のような最適化方法により、効率的に求められ得る。

【００２３】

マルコフ決定過程の線形公式は、以下に示されるように、制御コストを定義すべく、且つ、車両動力学的値に関する条件を加えるべく使用され得る。

【数５】

$$r(\mathbf{x}_k, \mathbf{u}_k) := q(\mathbf{x}_k) + KL(p(\mathbf{x}_{k+1}|\mathbf{x}_k) \| p(\mathbf{x}_{k+1}|\mathbf{x}_k, \mathbf{u}_k)) \quad (3)$$

$$p(\mathbf{x}_{k+1}|\mathbf{x}_k) = 0 \Longrightarrow \forall \mathbf{u}_k, p(\mathbf{x}_{k+1}|\mathbf{x}_k, \mathbf{u}_k) = 0 \quad (4)$$

ここで、

$q(\mathbf{x}) \geq 0$

は状態コスト関数であり、且つ、

$KL(\cdot\|\cdot)$

はクルバック－ライブラー（ＫＬ）偏差である。式（３）は、動作のコストを、それが系に対して有する確率論的効果の量に対して関連付け、且つ、それを状態コストに対して加算する。第２の条件は、何らの動作も、受動的動力学の下では達成され得ない新たな遷移を導入しないことを確実とする。式（１）により表された確率論的動力学系は、当然、上記仮定を満足する。

【００２４】

ハミルトン－ヤコビ－ベルマン方程式（式（２））は、Ｌ－ＭＤＰ形態において、指数的に変換された到達コスト関数に対する線形微分方程式（以下、線形化ベルマン方程式という）へと書き換えられ得る。

【数６】

$$Z_{avg} Z(\mathbf{x}_k) = \exp(-q(\mathbf{x}_k)\Delta t)\mathcal{G}[Z](\mathbf{x}_{k+1})$$

$$Z(\mathbf{x}) := \exp\left(-V^*(\mathbf{x})\right), \quad (5)$$

$$Z_{avg} := \exp\left(-V^*_{avg}\right)$$

$$p(\mathbf{x}_{k+1}|\mathbf{x}_k, \pi^*_k) = \frac{p(\mathbf{x}_{k+1}|\mathbf{x}_k)Z(\mathbf{x}_{k+1})}{\mathcal{G}[Z](\mathbf{x}_{k+1})}$$

式中、

$Z(\mathbf{x})$

及びZ_{avg}は、それぞれ、Ｚ値と称される指数的に変換された到達コスト関数、及び、最適ポリシーの下での平均コストである。（式（１））における状態遷移はガウス性であることから、制御された動力学的値と受動的な動力学的値との間のＫＬ偏差は、

【数７】

$$KL\left(p(\mathbf{x}_{k+1}|\mathbf{x}_k) \,\|\, p(\mathbf{x}_{k+1}|\mathbf{x}_k,\mathbf{u}_k)\right) = \frac{1}{2\rho(\mathbf{x}_k)}\mathbf{u}_k^{\top}\mathbf{u}_k$$

$$\frac{1}{\rho(\mathbf{x}_k)} := B(\mathbf{x}_k)^{\top}\left(C(\mathbf{x}_k)^{\top}C(\mathbf{x}_k)\right)^{-1}B(\mathbf{x}_k) \qquad (6)$$

として表される。

【００２５】

その後、Ｌ－ＭＤＰ系に対する最適な制御ポリシーは、

【数８】

$$\pi^*(\mathbf{x}_k) = -\rho(\mathbf{x}_k)B(\mathbf{x}_k)^{\top}V_{\mathbf{x}_k} \qquad (7)$$

として表され、式中、

$V_{\mathbf{x}_k}$

は、x_kにおけるxに関する到達コスト関数Ｖの偏微分値である。Ｚ値及び平均コストは、系の動力学的値が完全に入手可能であるとき、固有値又は固有関数を解くことにより、線形化ベルマン方程式から導かれ得る。

【００２６】

本明細書中に記述されたコンピュータ処理システム１４の実施形態は、種々の形式の入力及び出力情報を測定、受信、及び／又は、アクセスすることによりシステム（例えば、車両）の状態ｘ（ｔ）を決定する。例えば、データは、このシステムに結合されたセンサ又はこのシステムと別途通信するセンサを用いて測定され得る。コンピュータ処理システム１４は、制御入力ｕを決定することで、式（１）により特徴付けられる車両の安定性及び所望の運動を達成し且つ式（２）において記述されたエネルギに基づくコスト関数を最小化する。

【００２７】

本明細書中に記述されたコンピュータ処理システム１４の実施形態は、相互に作用する２つの学習システム又は学習ネットワーク、すなわちactorネットワーク（又は「actor」）８３及びcriticネットワーク（又は「critic」）８１を含む。これらネットワークは、人工ニューラルネットワーク（ＡＮＮ）を用いて実現され得る。actor８３は、状態依存の制御ポリシーを使用して、車両に対して適用され且つ到達コスト（cost-to-go）に対する最小値を生成する制御入力ｕ（ｘ）を決定する。actorは、平均コストと、近似的到達コスト関数から決定された推定到達コストと、車両の現在の状態に対する制御用動力学的値と、受動的に収集されたデータとを用いてノイズレベルを推定することにより、制御入力を決定すべく構成される。critic８１は、受動的に収集されたデータのサンプル及び状態コ

ストを用いて、推定平均コストと、actorネットワークにより適用されたときに車両の到達コストに対する最小値を生成する近似された到達コスト関数とを決定する。本明細書中に開示される幾つかの実施形態において、actor８３は内部ループフィードバックコントローラとして実現されると共に、critic８１は、外部ループフィードバックコントローラとして実現される。両者ともに、制御命令をもたらすように作動可能である車両を起動可能な機構又は制御器に関するフィードフォワード経路中に配置される。

【００２８】

受動的に収集されたデータのサンプルと、車両制御用動力学的モデル８７から受信された状態コストｑ（ｘ）とを用いて、critic８１は、車両の現在の状態x_k、次の状態x_{k+1}、及び、最適ポリシー下での状態コストq_kを評価し、且つ、先に記述されたベルマン方程式の線形化版（式（５））を使用して、近似された到達コスト関数

$\hat{Z}(\mathbf{x})$

（Ｚ値）を決定し、且つ、actor８３により使用される推定平均コスト

$\hat{Z}_{avg}$

を生成する。Ｚ値を推定するために、重み付けされた放射基底関数（ＲＢＦ）の線形結合(linear combination)が使用され得る。

【数９】

$$Z(\mathbf{x}) \approx \hat{Z}(\mathbf{x}) := \sum_{j=0}^{N} \omega_j f_j(\mathbf{x})$$

式中、ωは重みであり、f_jは第ｊ番目のＲＢＦであり、且つ、ＮはＲＢＦの個数である。基底関数は、車両システムの非線形の動力学に依存して適切に選択され得る。重みは、指数化された真の到達コストと推定到達コストとの間における最小二乗誤差を最小化することにより、最適化される。

$Z(\mathbf{x}_k)$

及び

Z_{avg}

を真のＺ値及び真の平均Ｚ値コストとし、且つ

$\hat{Z}_k$

、

$\hat{Z}_{avg}$

をそれぞれそれらの推定対応物とする。近似された到達コスト関数は、criticネットワークによりリアルタイムで学習され得る。

【数１０】

$$\min_{\omega,\hat{Z}_{avg}} \frac{1}{2}\sum_{D}\left(\hat{Z}_{avg}\hat{Z}_k - Z_{avg}Z_k\right)^2, \tag{8}$$

$$s.t\sum_{i=0}^{N}\omega_i = C, \forall i\ \omega_i \geq 0, \forall \mathbf{x}\ \hat{Z}_{avg}\hat{Z}(\mathbf{x}) \leq 1, \tag{9}$$

式中、Ｃは自明な解ω＝０への収束を回避すべく使用される一定値である。第２及び第３の制約は、式（５）から由来する

$$\forall \mathbf{x}, 0 < Z_{avg}Z(\mathbf{x}) \leq 1,$$

と、

$$\forall \mathbf{x}, q(\mathbf{x}) \geq 0.$$

とを満足するために必要とされる。

【００２９】

重みω及び平均コスト

$\hat{Z}_{avg}$

は、真の到達コストと推定到達コストとの間の誤差を、線形化ベルマン方程式（ＬＢＥ）（式（５））から以下のように決定された近似的な時間差的誤差e_kにより近似することにより、ラグランジュ緩和時間差（ＴＤ）学習に基づいて更新され得る。なぜなら、真の到達コスト及び真の平均コストは、ｐＡＣ方法に対して使用された情報によっては決定されないからである。

【数１１】

$$\hat{Z}_{avg}\hat{Z} - Z_{avg}Z_k \approx e_k := \hat{Z}_{avg}\hat{Z}_k - \exp(-q_k)\hat{Z}_{k+1}$$

$$\tilde{\omega}^{i+1} = \omega^i - \alpha_1^i e_k Z_{avg}\mathbf{f}_k \tag{10}$$

$$\omega^{i+1} = \tilde{\omega}^i + \lambda_1 + \sum_{l=0}^{N}\lambda_{2l}\delta_{jl} + \lambda_3\hat{Z}_{avg}\mathrm{f}_k \tag{11}$$

$$\hat{Z}_{avg}^{i+1} = \hat{Z}_{avg}^{i} - \alpha_2{}^i e_k\hat{Z}_k \tag{11A}$$

式中、

α_1^i

及び

α_2

は、学習率であり、且つ、e_kはＬ－ＭＤＰに対するＴＤ誤差である。δ_{ij}はディラックのデルタ関数を表している。下付き文字ｉは、反復の回数を表している。λ_1、λ_2、λ_3は、制約式（９）に対するラグランジュ乗数である。ωは、式（１０）による誤差を最小化すべく、且つ、式（１１）による制約を満足すべく更新される。反復とは、（criticに対する重みω及びactorに対するμのような）critic及びactorのパラメータの更新として定義され得る。これに加え、criticネットワークのパラメータの更新は、車両が運動しているときに実施され得る。本明細書中に記述された方法において、criticネットワークの更新の間に使用される唯一のデータは、受動的に収集されたデータである。

【００３０】

各乗数の値は、以下の方程式を解くことにより算出される。

【数１２】

$$\begin{bmatrix} \sum_j f_j & f_0 \cdots f_N & \mathbf{f}^T\mathbf{f} \\ 1 & 1 \cdots 0 & f_0 \\ \vdots & \ddots & \vdots \\ 1 & 0 \cdots 1 & f_N \\ N & 1 \cdots 1 & \sum_j f_j \end{bmatrix} \begin{bmatrix} \lambda_1^i \\ \lambda_2^i \\ \lambda_3^i \end{bmatrix} = \begin{bmatrix} c - \sum_j \tilde{\omega}^i \\ -\tilde{\omega}^i{}_0 \\ \vdots \\ -\tilde{\omega}^i{}_N \\ 1 - \hat{Z}_{avg}\hat{Z}_k \end{bmatrix}$$

【００３１】

幾つかの場合、制約の部分集合は有効でないことがあり得る。斯かる場合、これらの制約に対する乗数はゼロに設定され、且つ、残りの有効な制約に対する乗数が求められる。criticは、受動的動力学の下での状態遷移サンプル

$(\mathbf{x}_k \mathbf{x}_{k+1})$

及び状態コストq_kを用いて、各パラメータを更新する。重みω、推定Ｚ値

$\hat{Z}$

、及び、平均コスト

$\hat{Z}_{avg}$

は、車両が運動している間に、式（１０）～（１１Ａ）に従い、オンラインで更新され得る。

【００３２】

コンピュータ処理システムにおいて、criticネットワークに作用的に結合されたactor８３は、制御入力を決定し、到達コストに対する最小値を生成する車両に適用され得る。criticにより生成された推定到達コスト

$\hat{Z}$

及び推定平均コスト

$\hat{Z}_{avg}$

と、状態コストｑ（ｘ）と、車両制御用動力学的モデル８７から決定された現在の状態に対する制御用動力学的値情報Ｂ（ｘ）と、criticにより使用されて到達コスト関数

$\hat{Z}(\mathbf{x})$

を推定且つ推定平均コスト

$\hat{Z}_{avg}$

を生成すべく受動的に収集されたデータのサンプルとを用い、actor８３は制御入力を決定し得る。制御入力は、制御ポリシーπを修正すべく使用され得る。特定の実施形態において、ポリシーπは、上述の態様で、収束するまで反復的に修正され、その時点でそれは最適化されたと見做される。actorは、標準的なベルマン方程式を用い、且つ、能動的探索なしで、制御ポリシーを改善する。制御用動力学的値は、車両に対する既知の制御用動力学的モデルから決定され得る。

【００３３】

actor８３はまた、制御入力ｕ（ｘ）を各車両システムに対してリアルタイムで適用し、所望の操作（例えば、高速道路合流、車線変更）を自律的に実施もし得る。本明細書中に開示される幾つかの実施形態において、actor８３は、内部ループフィードバックコントローラにおいて具現され得ると共に、critic８１は、外部ループフィードバックコントローラにおいて具現され得る。両者ともに、車両を起動し得る制御器に関するフィードフォワード経路中に配置される。

【００３４】

actor８３は、criticからの評価値（例えば、

$\hat{Z}$

及び

$\hat{Z}_{avg}$

）、受動的動力学の下でのサンプル、及び、既知の制御用動力学的値を用いて、ノイズレベルρを推定することにより、制御ポリシーを改善又は修正する。ノイズレベルは、重み付けされた各放射基底関数の線形結合により、近似的に学習される。

【数１３】

$$\rho(\mathbf{x};\boldsymbol{\mu}) \approx \hat{\rho}(\mathbf{x};\boldsymbol{\mu}) := \sum_{j}^{M} \mu_j g_j(\mathbf{x}) \qquad (12)$$

式中、μｊは、ｊ番目の放射基底関数ｇｊに対する重みである。Ｍは、放射基底関数の個数である。

【００３５】

ρは、到達コストと動作－状態値との間の最小二乗誤差を最小化することにより、最適化される。

【数１４】

$$\min_{\mu}\frac{1}{2}\sum_{D}\left(\hat{Q}_k - V_k^* - V_{avg}^*\right)^2$$

式中、

V^*

、

V_{avg}^*

及び

$\hat{Q}$

は、最適な制御ポリシーの下での、真の到達コスト関数、平均コスト、及び推定された動作－状態値である。最適制御ポリシーは目的関数を最小化することにより学習され得る。なぜなら、真の動作－価値コストは、ポリシーが最適ポリシーであるとき且つそのときにのみ、

$V^* + V_{avg}^*$

に等しいからである。

$\hat{Z}_k$

及び

$\hat{Z}_{avg}$

は、ノイズレベルを更新するときに、以下の関係に従い、

$\hat{V}_k$

及び

$\hat{V}_{avg}$

を決定すべく使用され得る。

【数１５】

$$\hat{V}_k = -log(\hat{Z}_k)$$

$$\hat{V}_{avg} = -log(\hat{Z}_{avg})$$

【００３６】
重みμは、以下に定義される近似的な時間差的誤差ｄｋにより更新される。標準的なベルマン方程式は近似されて誤差ｄｋを決定する。なぜなら、真の到達コスト及び真の平均コストは算出されることができないからである。
【数１６】

$$\hat{Q}_k - V_k^* - V_{avg}^* \approx d_k := \hat{Q}_k - \hat{V}_k - \hat{V}_{avg}$$

$$\hat{Q}_k \approx q_k \Delta t + \frac{0.5\Delta t}{\hat{\rho}_k} \mathbf{u}_k^{\mathsf{T}} \mathbf{u}_k + \hat{V}(\mathbf{x}_{k+1} + B_k \mathbf{u}_k \Delta t)$$

$$\mathbf{u}_k \approx -\hat{\rho}_k B_k^{\mathsf{T}} \hat{V}_{\mathbf{k}},$$

式中、

$\mathbf{x}_{k+1}$

は、受動的動力学の下での次の状態であり、且つ、

$\mathbf{x}_{k+1} + B_k \mathbf{u}_k \Delta t$

は、動作ｕｋによる制御された動力学の下での次の状態である。推定到達コスト、平均コスト、及び、それらの微分値は、criticからの推定されたＺ値及び平均Ｚ値コストを利用することにより算出され得る。更に、

$\hat{V}(\mathbf{x}_{k+1} + B_k \mathbf{u}_k \Delta t)$

は、
【数１７】

$$\hat{V}(\mathbf{x}_{k+1} + B_k \mathbf{u}_k \Delta t) \approx \hat{V}_{k+1} + \hat{V}_{x_{k+1}}^{\mathsf{T}} B_k \mathbf{u}_k \Delta t.$$

により近似されて、
μ

に関してＴＤ誤差を線形化し得る。
【００３７】

μ
は、近似されたＴＤ誤差による時間差（ＴＤ）学習を用いて更新され得る。
【数１８】

$$\mu^{i+1} = \mu^{i} - \beta^{i} d_k L_{k,k+1} g_k,$$

$$d_k \approx q_k \Delta t - \hat{V}_{k+1} + \hat{V}_k + \hat{V}_{avg} + L_{k,k+1} \rho_k,$$

$$L_{k,k+1} := \left(0.5\hat{V}_{\mathbf{k}} - \hat{V}_{\mathrm{k+1}}\right)^{\mathrm{T}} B_k B_k^{\mathrm{T}} \hat{V}_k \Delta t$$

式中、β^{i}は学習率であり、且つ、$L_{k,k+1}$は、項Ｌ（x_k，x_{k+1}）の省略版である。
【００３８】
この手順は、与えられた状態において、受動的動力学的値、状態コストq_k、及び、制御用動力学的値B_kの下で、状態遷移サンプル
$(\mathbf{x}_k, \mathbf{x}_{k+1})$

を用いることにより、能動的探索なしでポリシーを改善する。標準的なactor－critic方法は、能動的探索によりポリシーを最適化する。定義されたこれらのactor及びcriticの機能により、コンピュータ処理システム１４は、Ｌ－ＭＤＰを用いて、半モデルフリー強化学習を実現し得る。
【００３９】
本明細書中に記述された方法において、ポリシーは、受動的に収集されたデータのサンプルと、車両制御用動力学の知見とを用い、到達コストと動作－状態値との間の誤差を最小化することにより学習されるパラメータにより最適化される。本明細書中に記述された方法は、乗用車を制御すべく通常的に利用可能である車両自体の動力学的モデルにより、最適ポリシーが決定されることを可能とする。上記方法はまた、それらの動力学的モデルが通常は既知でない周囲の車両の操作に関して受動的に収集されたデータも使用する。これに加え、本明細書中に記述された方法を用いると、最適な制御ポリシーを決定する上で、車両環境の受動的動力学的値Ａ（x_t）及び過渡的ノイズレベルＣ（x_t）は、認識される必要はない。
【００４０】
図２は、本明細書中に記述された方法に係る、コンピュータ処理システム１４における、制御入力の決定、及び、制御ポリシーの修正又は最適化の実行中の情報の流れを示す概略図である。従来のactor－critic方法は、周囲環境から能動的に収集されたデータのサンプルを用いて動作し得る一方、本明細書中に記述されたｐＡＣ方法は、周囲環境の能動的探索なしで、代わりに、受動的に収集されたサンプル、及び、既知の車両制御用動力学的モデルを用いて、最適な制御ポリシーを決定する。critic８１又はactor８３において受信された一切の情報は、後で使用するためにメモリ内にバファリングされ得る。例えば、パラメータ値を算出し又は推定すべくcritic又はactorに必要とされる情報の全てが現在は入手できないという状況において、受信情報は、残りの必要な情報が受信されるまで、バッファリングされ得る。
【００４１】
図３は、本明細書中に開示された幾つかの実施形態に従い、制御入力を決定し且つ制御ポリシーを修正及び／又は最適化するための図１のコンピュータ処理システムの動作を示すフローチャートである。

【００４２】
プロセスは、ブロック３１０にて開始され、そこでcritic８１は、推定平均コスト
$\hat{Z}_{avg}$

と、actorネットワークにより適用されたときに車両の到達コストに対する最小値を生成する近似された到達コスト関数
$\hat{Z}(x)$

とを決定し得る。
【００４３】
次に、ブロック３２０にて、actor８３は、到達コスト関数
$\hat{Z}(x)$

を用いて、車両に適用されて該車両の到達コストに対する最小値を生成する制御入力を決定し得る。actor８３は、制御ポリシーπを修正して、このポリシーを改善し且つ／又はこの制御ポリシーを最適化し得る。
【００４４】
ブロック３３０においては、ブロック３２０において導かれた制御入力が車両に適用されて、例えば高速道路への合流又は車線の変更などの、車両の自律的動作が行われ得る。また、車両は、任意の改善又は最適化された制御ポリシーπに従って更に制御され得る。特定の実施形態において、車両操作は、制御ポリシーが、未だ最適化されたと考えられる点まで改善されていないとしても、ポリシーの最新版を用いてコンピュータ処理システムにより制御され得る。
【００４５】
ブロック３４０においては、actor８３及びcritic８１の種々のパラメータが更新され得る。この更新は、本明細書中に記述された関係に従って実施され得る。この更新に対して使用される唯一のデータは、受動的に収集されたデータであり得る。或る実施形態において、actor及びcriticは、それらのそれぞれのパラメータの更新を実施し得る。或いは、actor及びcriticのパラメータの更新は、ポリシー反復器（図示せず）又は同様の手段により実施され得る。
【００４６】
図４は、図１のコンピュータ処理システム１４と同様の態様で構成されたコンピュータ処理システム１１４が組み込まれた例示的な実施形態に係る車両１１を示す機能的ブロック図である。車両１１は、乗用車、トラック、又は、本明細書中に記述された操作を実施し得る他の任意の車両の形態を取り得る。車両１１は、完全に又は部分的に自律モードで動作すべく構成され得る。自律モードで動作している間、車両１１は、人的相互作用なしで動作すべく構成され得る。例えば、高速道路の合流操作が実行されている自律モードにおいて、車両は、車両乗員からの入力なしで、高速道路上の車両から安全距離を維持すること、他の車両と速度を調和すること等を行うように、スロットル、ブレーキ及び他のシステムを動作させ得る。
【００４７】
車両１１は、コンピュータ処理システム１１４に加え、且つ、相互に作用的に通信する種々のシステム、サブシステム及び構成要素、及び構成要素、例えば、センサシステム又は配列２８、一つ以上の通信インタフェース１６、操舵システム１８、スロットルシステム２０、制動システム２２、電源３０、動力システム２６、並びに本明細書中に記述されたように車両を動作させるために必要な他のシステム及び構成要素を含み得る。車両１１は、図４に示されたよりも多い又は少ないサブシステムを含み得ると共に、各サブシステム

は、複数の要素を含み得る。更に、車両１１のサブシステム及び要素の各々は、相互接続され得る。車両１１の記述された機能及び／又は自律的動作の一つ以上の実施は、相互に協働して動作している複数の車両システム及び／又は構成要素により実行され得る。

【００４８】

センサシステム２８は、任意の適切な形式のセンサを含み得る。本明細書中には、異なる形式のセンサの種々の例が記述される。しかし、実施形態は、記述された特定のセンサに限定されないことは理解される。

【００４９】

センサシステム２８は、車両１１の外部環境に関する情報を検知すべく構成された所定数のセンサを含み得る。例えば、センサシステム２８は、全地球測位システム（ＧＰＳ）のようなナビゲーションユニット、及び、例えば、慣性測定装置（ＩＭＵ）（図示せず）、ＲＡＤＡＲユニット（図示せず）、レーザ測距計／ＬＩＤＡＲユニット（図示せず）、及び車両の内部及び／又は該車両１１の外部環境の複数の画像を捕捉すべく構成されたデバイスを備える一台以上のカメラ（図示せず）等の他のセンサを含み得る。カメラは、スチルカメラ又はビデオカメラであり得る。ＩＭＵは、慣性加速度に基づいて車両１１の位置及び向きの変化を検知するように構成されたセンサ（例えば、加速度計及びジャイロスコープ等）の任意の組合せを組み込み得る。例えば、ＩＭＵは、車両のロール速度、ヨーレート、ピッチ速度、長手方向加速度、横方向加速度、及び、垂直加速度のようなパラメータを検知し得る。ナビゲーションユニットは、車両１１の地理的位置を推定すべく構成された任意のセンサであり得る。この目的の為に、ナビゲーションユニットは、地球に対する車両１１の位置に関する情報を提供するように作動可能な送受信機を含む一つ以上の送受信機を含み得る。また、ナビゲーションユニットは、業界公知の態様で、記憶され且つ／又は利用可能な地図を用いて与えられた開始点（例えば、車両の現在位置）から、選択された目的地までの走行ルートを決定又は計画するように構成され得る。

【００５０】

公知の態様において、車両センサ２８は、種々の車両システムに対する適切な制御命令を策定且つ実行する際にコンピュータ処理システム１１４により使用されるデータを提供する。例えば、慣性センサ、車輪速度センサ、道路状態センサ、及び操舵角センサからのデータは、車両を旋回させるための命令を策定して操舵システム１８において実行する上で、処理され得る。各車両センサ２８は、車両１１に組み込まれる任意の運転者支援機能及び自律的動作機能をサポートするために必要とされる任意のセンサを含み得る。センサシステム２８が複数のセンサを含む構成において、センサは、相互から独立的に作動し得る。代替的に、各センサのうちの２つ以上が、相互に協働して作動し得る。センサシステム２８のセンサは、コンピュータ処理システム１４に対し、及び／又は車両１１の他の任意の要素に対し、作用的に接続され得る。

【００５１】

また、各車両センサ２８により収集された任意のデータは、本明細書中に記述された目的でデータを必要とし又は利用する任意の車両システム又は構成要素にも送信され得る。例えば、車両センサ２８により収集されたデータは、コンピュータ処理システム１１４に、又は一つ以上の専用のシステム又は構成要素のコントローラ（図示せず）に送信され得る。付加的な特定の形式のセンサとしては、本明細書中に記述された機能及び操作を実施するために必要とされる他の任意の形式のセンサが挙げられる。

【００５２】

特定の車両センサからの情報は、一つよりも多い車両システム又は構成要素を制御すべく処理かつ使用され得る。例えば、自動化された操舵制御及び制動制御の両方を組み込んだ車両において、種々の道路状態センサは、データをコンピュータ処理システム１１４に提供し、このコンピュータ処理システムは、プロセッサが実行可能な記憶された命令に従って道路状態情報を処理すると共に、操舵システム及び制動システムの両方に対して適切な制御命令を策定することができるようになる。

【００５３】

車両１１は、センサの出力信号又は他の信号が、コンピュータ処理システム１１４又は別の車両システム若しくは要素による使用の前に前処理を必要とするという状況、又はコンピュータ処理システムから送信された制御信号が、起動可能なサブシステム又はサブシステム構成要素（例えば、操舵システム又はスロットルシステムの構成要素）による使用の前に処理を必要とするという状況に適した、信号処理手段３８を含み得る。信号処理手段は、例えば、アナログ／デジタル（Ａ／Ｄ）変換器又はデジタル／アナログ（Ｄ／Ａ）変換器であり得る。

【００５４】

センサ統合機能（sensor fusion capability）１３８は、センサシステム２８からのデータを入力として受け入れるべく構成されたアルゴリズム（又は、アルゴリズムを記憶するコンピュータプログラム製品）の形態であり得る。上記データは、例えば、センサシステム２８の各センサにて検知された情報を表すデータを含む。センサ統合アルゴリズムは、センサシステムから受信したデータを処理し、（例えば、複数の個別的なセンサの出力から形成された）統合された又は合成された信号を生成し得る。センサ統合アルゴリズム１３８は、例えば、カルマンフィルタ、ベイジアンネットワーク、又は、別のアルゴリズムを含む。センサ統合アルゴリズム１３８は更に、センサシステム２８からのデータに基づく種々のアセスメントを提供し得る。例示的な実施形態において、アセスメントは、車両１１の環境における個別的な物体又は特定構造の評価、特定状況の評価、及び、特定の状況に基づく可能的な影響の評価を含み得る。他のアセスメントも可能である。センサ統合アルゴリズム１３８は、コンピュータ処理システム１１４に組み込まれた又はコンピュータ処理システム１１４と作用的に通信する（メモリ５４のような）メモリ内に記憶され得ると共に、当業界において公知の態様でコンピュータ処理システムにより実行され得る。

【００５５】

本明細書中に記述された任意の情報若しくはパラメータの受信、収集、監視、処理、及び／又は、決定を参照するときにおける「連続的に」という語句の使用は、コンピュータ処理システム１１４が、これらのパラメータに関する情報が存在し又は検出されるや否や、又は、センサの取得サイクル及びプロセッサの処理サイクルに従ってできるだけ素早く、任意の情報を受信及び／又は処理すべく構成されることを意味している。コンピュータ処理システム１１４が、例えば、センサからのデータ又は車両構成要素の状況に関する情報を受信すると直ちに、コンピュータ処理システムは、記憶されたプログラム命令に従って動作し得る。同様に、コンピュータ処理システム１１４は、センサシステム２８から及び他の情報源から、同時進行的又は連続的に情報の流れを受信して処理し得る。この情報は、本明細書中に記述された態様及び目的にて、メモリ内に記憶された命令に従って処理及び／又は評価される。

【００５６】

また、図４は、先に記述されたように、図１のコンピュータ処理システム１４と同様の態様で構成された代表的なコンピュータ処理システム１１４のブロック図も示している。本明細書中に記述されたようにポリシーの修正を実施すると共に制御入力を決定するために必要とされる機能を組み込むと共に、コンピュータ処理システム１１４は、他の車両システム及び要素に作用的に接続されると共に、その他の点では、車両１１及びその構成要素の制御及び動作に影響するように構成され得る。コンピュータ処理システム１１４は、少なくとも幾つかのシステム及び／又は構成要素を、（ユーザ入力なしで）自律的に且つ／又は（一定程度のユーザ入力を以て）半自律的に制御すべく構成され得る。また、コンピュータ処理システムは、幾つかの機能を自律的及び／又は半自律的に制御及び／又は実行するようにも構成され得る。コンピュータ処理システム１１４は、種々のサブシステム（例えば、動力システム２６、センサシステム２８、操舵システム１８）から、各通信インタフェース１６のうちの任意のものから、及び／又は他の任意で適切な情報源から受信した入力及び／又は情報に基づき、車両１１の機能性を制御し得る。

【００５７】

図４の実施形態において、コンピュータ処理システム１１４は、図１に関して先に記述さ

れたように、車両制御用動力学的モデル１８７、critic１８１、actor１８３、及び、制御ポリシー２０１を含み得る。コンピュータ処理システム１１４は、先に記述されたように、制御入力を決定すべく、且つ自律車両の操作制御ポリシーを修正及び／又は最適化すべく構成され得る。また、コンピュータ処理システム１１４は、制御入力に従って、且つ、本明細書中に記述されたように修正又は最適化された制御ポリシーにも従って、車両を制御して所望操作を実施すべく構成され得る。

【００５８】

コンピュータ処理システム１１４は、図４に示された要素の幾つか又は全てを有し得る。加えて、コンピュータ処理システム１１４は、特定の用途に必要とされ又は所望される付加的な構成要素も含み得る。また、コンピュータ処理システム１１４は、複数のコントローラ又はコンピュータ処理デバイスであって、分散態様にて、情報を処理し且つ／又は車両１１の個別的な構成要素若しくはサブシステムを制御するように機能する複数のコントローラ又はコンピュータ処理デバイスを表し、又は、それにより具現され得る。

【００５９】

メモリ５４は、単一又は複数のプロセッサ５８により実行されて、図１に関して上述されたものを含む、車両１１の種々の機能を実行するデータ６０及び／又は命令５６（例えば、プログラムロジック）を収納し得る。メモリ５４は、本明細書中に記述された車両システム及び／又は構成要素（例えば、動力システム２６、センサシステム２８、コンピュータ処理システム１１４、及び、通信インタフェース１６）のうちの一つ以上にデータを送信し、それらからデータを受信し、それらと相互作用し、又はそれらを制御するための命令を含む、付加的な命令も含み得る。命令５６に加え、メモリ５４は、他の情報の中でも、道路地図、経路情報のようなデータを記憶し得る。斯かる情報は、自律的、半自律的、及び／又は手動的なモードにおける車両１１の動作の間において、ルートを計画するのに且つその他にことをするのに、車両１１及びコンピュータ処理システム１１４により使用され得る。

【００６０】

コンピュータ処理システム１１４は、（概略的に６２と表される）一つ以上の自律的な機能又は動作を実施するために、種々の起動可能な車両システム及び構成要素の制御を連携調整するように構成され得る。これらの自律的な機能６２は、メモリ５４及び／又は他のメモリ内に記憶されると共に、プロセッサにより実行されたときに、本明細書中に記述された種々のプロセス、命令又は機能のうちの一つ以上を実現するコンピュータ可読プログラムコードの形態で実現され得る。

【００６１】

通信インタフェース１６は、車両１１と、外部センサ、他の車両、他のコンピュータシステム、（本明細書中に記述されたように、衛星システム、携帯電話／無線通信システム、種々の車両サービスセンターなどのような）種々の外部のメッセージ及び通信システム、及び／又はユーザとの間の相互作用を許容すべく構成され得る。通信インタフェース１６は、車両１１のユーザに情報を提供し又はユーザから入力を受信するためのユーザインタフェース（例えば、一台以上のディスプレイ（図示せず）、音声／オーディオインタフェース（図示せず）、及び／又は他のインタフェース）を含み得る。

【００６２】

また、通信インタフェース１６は、ワイドエリアネットワーク（ＷＡＮ）、無線通信ネットワーク、及び／又は他の任意で適切な通信ネットワークにおける通信を可能とするインタフェースも含み得る。通信ネットワークは、有線の通信リンク、及び／又は無線の通信リンクを含み得る。通信ネットワークは、上記のネットワーク及び／又は他の形式のネットワークの任意の組合せを含み得る。通信ネットワークは、一つ以上のルータ、スィッチ、アクセスポイント、無線アクセスポイント、及び／又は類似物を含み得る。一つ以上の構成において、通信ネットワークは、任意の近傍車両及び車両１１と、任意の近傍の路側の通信ノード及び／又はインフラとの間の通信を許容し得る、車両対全て（Ｖ２Ｘ）（車両対インフラストラクチャ（Ｖ２Ｉ）技術及び車両対車両（Ｖ２Ｖ）技術を含む）の技術

を包含し得る。
【００６３】
ＷＡＮネットワーク環境において使用されたとき、コンピュータ処理システム１１４は、ネットワーク（例えば、インターネット）のようなＷＡＮ上での通信を確立するためのモデム又は他の手段を含み（又は、それに対して作用的に接続され）得る。無線通信ネットワークにおいて使用されたとき、コンピュータ処理システム１１４は、無線ネットワークにおける一つ以上のネットワークデバイス（例えば、基地送受信ステーション）を介して無線コンピュータ処理デバイス（図示せず）と通信するための一つ以上の送受信機、デジタル信号プロセッサ、及び付加的な回路機構並びにソフトウェアを含み（又は、それに対して作用的に接続され）得る。これらの構成は、種々の外部情報源から定常的な情報の流れを受信する種々の態様を提供する。
【００６４】
車両１１は、コンピュータ処理システム１１４並びに他の車両システム及び／又は構成要素と作用的に通信し且つコンピュータ処理システムから受信した制御命令に応じて作用し得る、種々の起動可能なサブシステム及び要素を含み得る。種々の起動可能なサブシステム及び要素は、（例えば、ＡＣＣ及び／又は車線維持などの）いずれの自律的の走行支援システムが起動されているのか且つ／又は車両が完全自律モードで駆動されているのかといった所定の走行状況のような要因に依存して、手動的又は（コンピュータ処理システム１１４により）自動的に制御され得る。
【００６５】
操舵システム１８は、車両ホイール、ラック及びピニオン操舵ギア、操舵ナックル、及び／若しくは車両１１の方向を調節すべく作用可能であり得る他の任意の要素（コンピュータシステムで制御可能な任意の機構又は要素を含む）、又は要素の組み合わせを含み得る。動力システム２６は、車両１１に動力運動を提供すべく作用可能な構成要素を含み得る。例示的な実施形態において、動力システム２６は、エンジン（図示せず）、（ガソリン、ディーゼル燃料、又は、ハイブリッド車両の場合には一つ以上の電気バッテリのような）エネルギ源、及び、変速機（図示せず）を含み得る。制動システム２２は、車両１１を減速すべく構成された、要素及び／又はコンピュータシステムで制御可能な任意の機構の任意の組合せを含み得る。スロットルシステムは、（例えば、加速ペダル、及び／又は例えばエンジンの作動速度を制御することで車両１１の速度を制御するように構成された任意のコンピュータシステム制御可能な機構などの）要素及び／又は機構を含み得る。図１は、車両に組み込まれ得る車両サブシステムの僅かな例１８、２０、２２、２６を示している。特定の車両は、これらのシステムの一つ以上、又は示されたシステムの一つ以上に加えて他のシステム（図示せず）の一つ以上を組み込み得る。
【００６６】
車両１１は、コンピュータ処理システム１１４、センサシステム２８、起動可能なサブシステム１８、２０、２２、２６、及び他のシステム並びにこれらの要素が、コントローラエリアネットワーク（ＣＡＮ）バス３３又はその類似物を用いて相互に通信し得るように構成され得る。ＣＡＮバス及び／又は他の有線又は無線の機構を介し、コンピュータ処理システム１４は、種々の車両システム及び構成要素に対してメッセージを送信し（且つ／又は、それらからメッセージを受信し）得る。或いは、本明細書中に記述された要素及び／又はシステムの任意のものは、バスを使用せずに相互に対して直接的に接続され得る。同様に、本明細書中に記述された要素及び／又はシステム間の接続は、（有線接続のような）別の物理的媒体を経由され得るか、又は上記接続は無線接続であり得る。
【００６７】
図１は、コンピュータ処理システム１４、メモリ５４、及び通信インタフェース１６のような車両１１の種々の構成要素を、車両１１に一体化されているとして示しているが、これら構成要素の一つ以上は、車両１１とは別体的に取付けられ、又は関連付けられ得る。例えば、メモリ５４は、部分的に又は完全に、車両１１とは別体的に存在し得る。したがって、車両１１は、別体的又は一体的に配置され得る複数のデバイス要素の形態で提供さ

れ得る。車両１１を構成するデバイス要素は、有線又は無線の態様で相互に通信的に結合され得る。

【００６８】

実施例

図５及び図６を参照すると、本明細書中に記述された制御入力及びポリシー修正／最適化方法の実施形態の一つの実施例において、自律的な高速道路合流操作がシミュレートされる。この操作は、４次元の状態空間、及び、１次元の動作空間を有する。動力学的値は

【数１９】

$$\mathbf{x} = [dx_{12}, dv_{12}, dx_{02}, dv_{02}]^{\mathsf{T}},$$
$$\mathrm{A}(\mathbf{x}) = [dx_{12}, 0, dv_{02}, +0.5\alpha_0(\mathbf{x})\Delta t, \alpha_0(\mathbf{x})]^{\mathsf{T}}$$
$$B(\mathbf{x}) = [0.5\Delta t, 1, 0, 0]^{\mathsf{T}}, \mathrm{C}(\mathbf{x}) = [0, 2.5, 0, 2.5]^{\mathsf{T}}$$
$$\alpha_0(\mathbf{x}) = \alpha \frac{v_2^{\beta}(-dv_{02})}{-dx_{02}^{\gamma}}, \Delta t = 0.1[\mathrm{sec}]$$

であり、式中、下付き文字０は、高速道路の最右側車線上で合流車両の後方の車両（「後続車両」という）を表し、１は、ランプＲＲ上で合流している自動化車両を表し、且つ、２は、高速道路上の最右側車線上で合流車両の前方の車両（「先行車両」という）を表している。dx_{12}及びdv_{12}は、先行車両からの合流車両の相対的な位置及び速度を表している。例示な目的で、先行車両は一定速度v_2＝３０メートル／秒で走行されること、及び、後続車両に対する車両制御用動力学的モデルは既知であることが仮定される。もし後続車両の速度が先行車両よりも低速である（dv_{02}＜０）場合には、α＝１．５５、β＝１．０８、γ＝１．６５であり、その他の場合には、α＝２．１５、β＝－１．６５、γ＝－０．８９である。状態コスト

$q(\mathbf{x})$

は、

【数２０】

$$q(\mathbf{x}) = k_1\left(1.0 - \exp\left(-k_2\left(1 - \frac{2dx_{12}}{dx_{02}}\right)^2 - k_3(dv_{10})^2\right)\right)$$

であり、式中、k_1、k_2及びk_3は、状態コストに対する重みである。もし合流車両がランプ上で後続車両と先行車両との間である（すなわち、dx_{12}＜０、及びdx_{12}＞dx_{02}という条件にある）なら、k_1＝１、k_2＝１０、及びk_3＝１０であり、さもなければ、k_1＝１０、k_2＝１０、及びk_3＝０である。コストは、自動車が、後続車両と先行車両との中間に、後続車両と同一の速度で合流することを誘起すべく設計される。初期状態は、－１００＜dx_{12}＜１００メートル、－１０＜dv_{12}＜１０メートル／秒、－１００＜dx_{02}＜－５メートル、及び－１０＜dv_{02}＜１０メートル／秒において、ランダムに選択される。Ｚ値を近似するのに、ガウス放射基底関数が使用された。

【数２１】

$$fi(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x}-m_i)^{\mathrm{T}} Si(\mathbf{x}-m_i)\right)$$

式中、ｍｉ及びＳｉは、第ｉ番目の放射基底関数に対する平均及び逆共分散である。高速道路合流のシミュレーションに対し、Ｚ値は、４,０９６個のガウス放射基底関数であって、それらの平均が、状態の次元毎に８個の値から成る格子の頂点上に設定されたガウス放射基底関数により近似された。各基底の標準偏差は、各次元において最も近い２つの基底間の距離の０.７であった。上記例において

$\rho(x)$

の実際の値は一定であるため、

$\rho(x)$

を推定するのにｇ（ｘ）＝１が使用された。上記方法は、受動的動力学的値をシミュレートすることにより収集された１０,０００個のサンプルから、ポリシーを最適化した。図６は、本明細書中に記述された方法により決定された順次的な制御入力を用い、１２５個の異なる初期状態から開始し、（収束に対して必要とされる反復の回数として表現された）３０秒以内に好首尾に合流する割合を示している。

【００６９】

上記の詳細な説明においては、その一部を構成する添付図面に対する参照が為されている。図において、同様の記号は典型的に、状況が別様に示唆するのでなければ、同様の構成要素を特定している。詳細な説明、図、及び、請求項中に記述された代表的実施形態は、限定的であることを意味しない。本明細書中に呈示された主題の有効範囲から逸脱せずに、他の実施形態が利用され得ると共に、他の変更が為され得る。概略的に本明細書中に記述されると共に各図中に示された本開示の各見地は、全てが本明細書において明示的に企図された多様な異なる構成にて、配置、置換、結合、分離、及び、設計され得ることは容易に理解される。

【００７０】

本開示を読んだに当業者により理解され得るように、本明細書中に記述された種々の見地は、方法、コンピュータシステム、又は、コンピュータプログラム製品として具現され得る。従って、それらの見地は、完全にハードウェアの実施形態、完全にソフトウェアの実施形態、又は、ソフトウェア及びハードウェアの見地を組み合わせた実施形態の形態を取り得る。更に、斯かる見地は、一種類以上のコンピュータ可読記憶媒体であって、本明細書中に記述された機能を実行するために当該記憶媒体内に又は当該記憶媒体上に具現されたコンピュータ可読プログラムコード又は命令を有するコンピュータ可読記憶媒体により記憶されたコンピュータプログラム製品の形態を取り得る。これに加え、本明細書中に記述されたデータ、命令又は事象を表す種々の信号は、送信元と送信先との間にて、金属ワイヤ、光ファイバ、及び／又は、（例えば、空気及び／又は空間などの）無線送信媒体のような信号導通媒体を通して進行する電磁波の形態で伝達され得る。

【００７１】

本明細書中で用いられるように、「一つの（a）」及び「一つの（an）」という語句は、一つ、又は、一つより多いものとして定義される。本明細書中で用いられるように、「複数の」という語句は、２つ、又は、２つより多いものとして定義される。本明細書中で用いられるように、「別の（another）」という語句は、少なくとも第２のもの、又は、そ

れより多いものとして定義される。本明細書中で用いられるように、「含む」及び／又は「有する」という語句は、備える（すなわち非制限的表現）として定義される。本明細書中で用いられるように、「～及び～の少なくとも一つ」という語句は、関連して列挙された対象物のうちの一つ以上の対象物の任意の全ての可能的な組み合わせを参照かつ包含する。一例として、「Ａ、Ｂ及びＣの少なくとも一つ」という表現は、Ａのみ、Ｂのみ、Ｃのみ、又は、（例えば、ＡＢ、ＡＣ、ＢＣ又はＡＢＣなどの）それらの任意の組合せを包含する。

【００７２】

従って、本発明の有効範囲を表すものとしては、上述の明細書ではなく、以下の各請求項に対して参照が為されるべきである。

本明細書に開示される発明は以下の実施態様を含む。

（１）車両の自律的動作を適応的に制御するコンピュータ実行型方法であって、該方法は、

ａ）車両を自律的に制御するように構成されたコンピュータ処理システムにおけるcriticネットワークにおいて、受動的に収集されたデータのサンプルと、状態コストとを用いて、推定平均コストと、actorネットワークにより適用されたときに車両の到達コストに対する最小値を生成する近似された到達コスト関数とを決定することと、

ｂ）コンピュータ処理システム内においてcriticネットワークに対して作用的に連結されたactorネットワークにおいて、車両に対して適用されて到達コストに対する最小値を生成する制御入力を決定すること、とを備え、

前記actorネットワークは、推定平均コストと、近似された到達コスト関数から決定された推定到達コストと、車両の現在の状態に対する制御用動力学的値と、受動的に収集されたデータのサンプルとを用いて、ノイズレベルを推定することにより制御入力を決定するように構成される、方法。

（２）前記近似された到達コスト関数は、以下の関係に従い、重み付けされた放射基底関数の線形結合を用いて決定され、

【数２２】

$$\hat{Z}(x) := \sum_{j=0}^{N} \omega_j f_j(x)$$

式中、ωは重みであり、f_jは第ｊ番目の放射基底関数であり、Ｎは近似された到達コスト関数を決定するために使用される放射基底関数の個数であり、且つ、

$\hat{Z}(x)$

は近似された到達コスト関数である、上記（１）に記載の方法。

（３）前記近似された到達コスト関数において使用される重みωは、以下の関係に従って更新され、

【数２３】

$$\omega^{i+1} = \tilde{\omega}^i + \lambda_1 + \sum_{l=0}^{N} \lambda_{2l}\delta_{jl} + \lambda_3 \hat{Z}_{avg} f_k$$

式中、δ_{jj}はディラックのデルタ関数を表し、上付き文字ｉは反復の回数を表し、λ_1、λ_2、λ_3はラグランジュ乗数であり、且つ

$\hat{Z}_{avg}$

は推定平均コストである、上記（２）に記載の方法。

（４）ベルマン方程式の線形化版を用いて決定された近似的な時間差的誤差を用いて前記criticネットワークのパラメータを更新する段階を更に備える、上記（１）に記載の方法。

（５）前記criticネットワークパラメータを更新する段階は、前記車両が運動しているときに実施される、上記（４）に記載の方法。

（６）前記criticネットワークにより決定された推定平均コストは、以下の関係に従って更新され、

【数２４】

$$\hat{Z}_{avg}^{i+1} = \hat{Z}_{avg}^{i} - \alpha_2{}^{i} e_k \hat{Z}_k$$

式中、βは学習率であり、e_kは近似的な時間差的誤差であり、

$\hat{Z}_k$

は前記近似された到達コスト関数から決定された推定コストであり、

$\hat{Z}_{avg}^{i}$

は状態ｉにおける推定平均コストであり、且つ、

$\hat{Z}_{avg}^{i+1}$

は状態ｉ＋１における推定平均コストである、上記（４）に記載の方法。

（７）前記受動的に収集されたデータは、前記criticネットワークパラメータを更新する間に使用される唯一のデータである、上記（４）に記載の方法。

（８）以下の関係に従って決定された近似的な時間差的誤差を用いて前記criticネットワークのパラメータを更新する段階を更に備え、

【数２５】

$$e_k := \hat{Z}_{avg}\hat{Z}_k - \exp(-q_k)\hat{Z}_{k+1}$$

式中、e_kは近似的な時間差的誤差であり、

$\hat{Z}_{avg}$

は推定平均コストであり、

$\hat{Z}_k$

は状態ｋにおける推定到達コストであり、

$\hat{V}_{k+1}$

は状態ｋ＋１における推定到達コストであり、且つ、q_kは状態ｋにおける状態コストである、上記（１）に記載の方法。

（９）前記近似された到達コスト関数は、前記criticネットワークにおいてリアルタイムで学習される、上記（１）に記載の方法。

（１０）前記ノイズレベルは、以下の関係に従い、重み付けされた基底関数の線形結合を用いて学習され、

【数２６】

$$\rho(x;\mu) \approx \hat{\rho}(x;\mu) := \sum_{j}^{M} \mu_j g_j(x)$$

式中、ρは推定ノイズレベルであり、μ_jは、g_jにより表された第ｊ番目の放射基底関数に対する重みであり、且つ、Ｍは、ノイズレベルを推定するために使用されるべき放射基底関数の個数である、上記（１）に記載の方法。

（１１）以下の関係に従って決定された近似誤差を用いて前記actorネットワークの重み付けパラメータを更新する段階を更に備え、

【数２７】

$$d_k \approx q_k \Delta t - \hat{V}_{k+1} + \hat{V}_k + \hat{V}_{avg} + L_{k,k+1}\rho_k,$$

式中、d_kは近似誤差であり、q_kは状態ｋにおける状態コストであり、

$\hat{V}_k$

は状態ｋにおいて近似された到達コストであり、

$\hat{V}_{k+1}$

は状態ｋ＋１において近似された到達コストであり、

$\hat{V}_{avg}$

は近似された平均コストであり、且つ、

【数２８】

$$L_{k,k+1} := \left(0.5\hat{V}_k - \hat{V}_{k+1}\right)^T B_k B_k^T \hat{V}_k \Delta t$$

であり、式中、B_kは状態ｋにおける制御用動力学的値である、上記（１０）に記載の方法

。
（１２）前記actorネットワークの重み付けパラメータの更新は、前記車両が運動しているときに実施される、上記（１１）に記載の方法。
（１３）前記actorネットワークの重み付けパラメータは、以下の関係に従って更新され、
【数２９】

$$\mu^{i+1} = \mu^{i} - \beta^{i} d_k L_{k,k+1} g_k,$$

式中、
μ^{i+1}
は状態ｉ＋１における重み付けパラメータの値であり、
μ^{i}
は状態ｉにおける重み付けパラメータの値であり、β^{i}は学習率であり、d_kは時間差的誤差であり、且つｇは放射基底関数である、上記（１１）に記載の方法。
（１４）　記制御入力を用い、前記自律的動作を制御すべく使用可能な制御ポリシーを修正する段階を更に備える、上記（１）に記載の方法。
（１５）前記推定平均コストが収束するまで、前記段階（ａ）及び（ｂ）を反復的に実施して前記制御入力を再決定することにより、前記自律的動作を制御するために使用可能な制御ポリシーを最適化する段階を更に備える、上記（１）に記載の方法。
（１６）前記制御ポリシーは、能動的探索なしで最適化される、上記（１５）に記載の方法。
（１７）車両の自律的動作を適応的に制御するように構成されたコンピュータ処理システムであって、該コンピュータ処理システムは、該コンピュータ処理システムの動作を制御する一つ以上のプロセッサと、該一つ以上のプロセッサにより使用可能なデータ及びプログラム命令を記憶するメモリとを備え、
前記一つ以上のプロセッサは、前記メモリ内に記憶された命令を実行して、
ａ）受動的に収集されたデータのサンプルと、状態コストとを用いて、推定平均コストと、前記車両の到達コストに対する最小値を生成する近似された到達コスト関数とを決定し、且つ、
ｂ）前記車両に対して適用されて前記到達コストに対する最小値を生成する制御入力を決定する、ように構成され、
前記一つ以上のプロセッサは、前記推定平均コストと、前記近似された到達コスト関数から決定された到達コストと、前記車両の現在の状態に対する制御用動力学的値と、受動的に収集されたデータのサンプルとを用いて、ノイズレベルを推定することにより制御入力を決定するように構成される、コンピュータ処理システム。
（１８）前記一つ以上のプロセッサは、前記メモリ内に記憶された命令を実行し、前記推定平均コストが収束するまで、前記段階（ａ）及び（ｂ）を反復的に実施して前記制御入力を再決定することにより、前記自律的動作を制御するために使用可能な制御ポリシーを最適化するように構成される、上記（１７）に記載のコンピュータ処理システム。
（１９）コンピュータシステムにより実行可能な命令が自身内に記憶された、一時的でないコンピュータ可読媒体であって、
ａ）受動的に収集されたデータのサンプルと、状態コストとを用いて、推定平均コスト

と、車両の到達コストに対する最小値を生成する近似された到達コスト関数とを決定することと、

ｂ）前記車両に対して適用されて前記到達コストに対する最小値を生成する制御入力を決定すること、とを備える機能を実施させ、

前記制御入力は、前記到達コストに対する最小値を生成し、且つ前記平均コストと、前記近似された到達コスト関数から決定された到達コストと、前記車両の現在の状態に対する制御用動力学的値と、受動的に収集されたデータのサンプルとを用いて、ノイズレベルを推定することにより制御入力が決定される、一時的でないコンピュータ可読媒体。

（２０）前記命令は、前記推定平均コストが収束するまで、前記段階（ａ）及び（ｂ）を反復的に繰り返して前記制御入力を再決定することにより、前記自律的動作を制御するために使用可能な制御ポリシーを最適化するように実行可能である、上記（１９）に記載の一時的でないコンピュータ可読媒体。

【図面】

【図１】

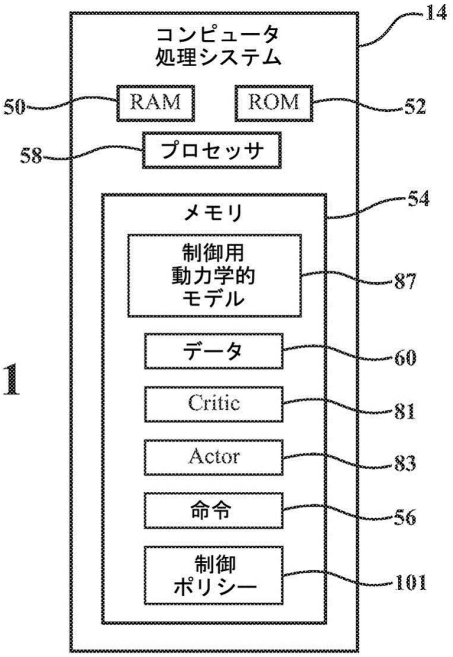


FIG. 1

【図２】

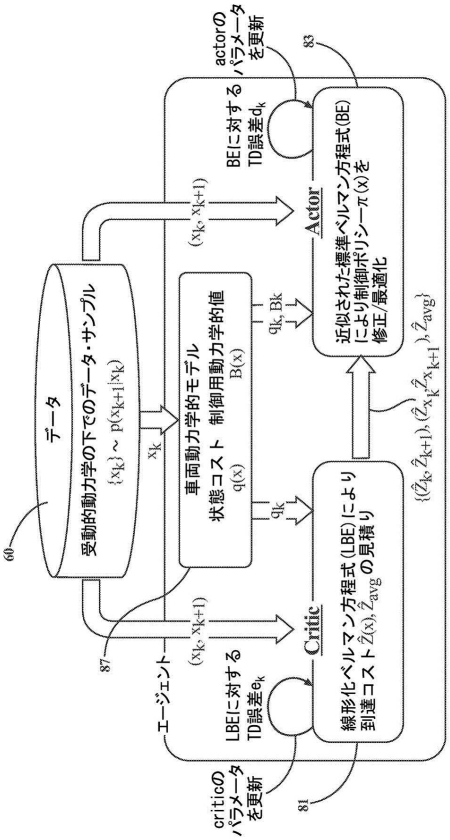
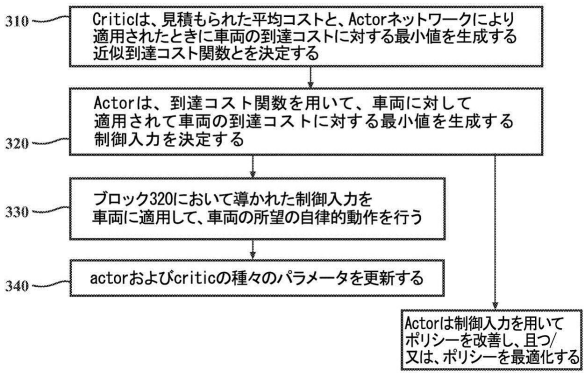


FIG. 2

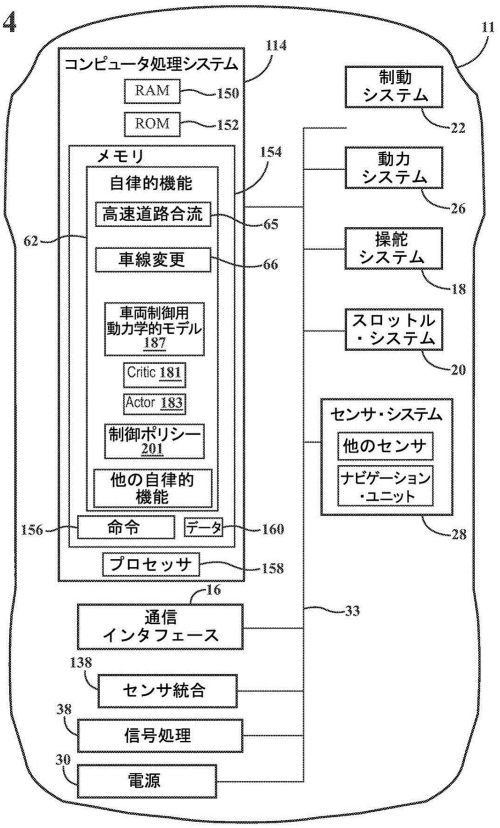
【図３】

FIG. 3



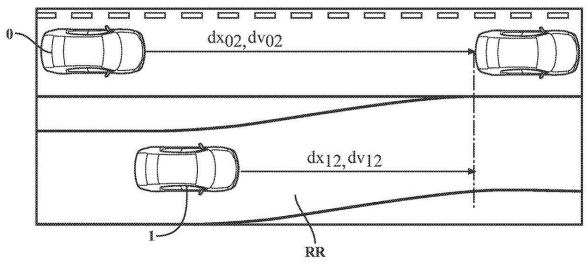
【図４】

FIG. 4



【図５】

FIG. 5



【図６】

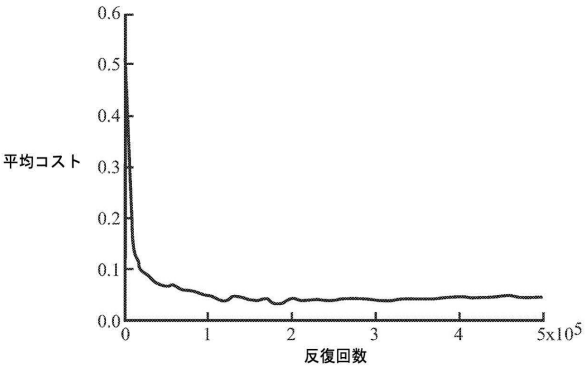


FIG. 6

フロントページの続き

(74)代理人　100123593
弁理士　関根 宣夫
(72)発明者　西 智樹
アメリカ合衆国，ケンタッキー ４１０１８，アーランガー，アトランティック アベニュ ２５，シー／オー トヨタ モーター エンジニアリング アンド マニュファクチャリング ノース アメリカ，インコーポレイティド
審査官　杉山 悟史

(56)参考文献
特開平１０－２５４５０５（ＪＰ，Ａ）
特開２００６－３１３５１２（ＪＰ，Ａ）
米国特許出願公開第２０１３／０２６２３５３（ＵＳ，Ａ１）
米国特許出願公開第２０１６／００９２７６４（ＵＳ，Ａ１）

(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０５Ｂ　１／００ －　７／０４
１１／００ －　１３／０４
１７／００ －　１７／０２
２１／００ －　２１／０２