(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7623879号
(P7623879)

(45)発行日　令和7年1月29日(2025.1.29)

(24)登録日　令和7年1月21日(2025.1.21)

(51)国際特許分類
Ｇ０６Ｎ　20/00　(2019.01)

ＦＩ
Ｇ０６Ｎ　20/00

請求項の数　6　（全40頁）

(21)出願番号　特願2021-67848(P2021-67848)
(22)出願日　令和3年4月13日(2021.4.13)
(65)公開番号　特開2022-162824(P2022-162824A)
(43)公開日　令和4年10月25日(2022.10.25)
審査請求日　令和6年2月22日(2024.2.22)

(73)特許権者　000005108
株式会社日立製作所
東京都千代田区丸の内一丁目６番６号
(74)代理人　110001678
藤央弁理士法人
(72)発明者　森澤 利浩
東京都千代田区丸の内一丁目６番６号
株式会社日立製作所内
(72)発明者　古林 雅佳
東京都千代田区丸の内一丁目６番６号
株式会社日立製作所内
審査官　千葉 久博

最終頁に続く

(54)【発明の名称】　行動制御計画装置、行動制御計画方法及び行動制御計画システム

(57)【特許請求の範囲】

【請求項１】

プロセッサとメモリを有して、他移動体の行動に対して自移動体の行動を計画する行動制御計画装置であって、

前記自移動体と他移動体の行動をシミュレーション条件として設定するシミュレーション条件設定部と、

前記シミュレーション条件と予め設定された機械学習モデルに基づいて、前記自移動体と他移動体の位置関係を所定の時間間隔毎にステップデータとして出力するシミュレータと、

前記ステップデータを取得して、前記機械学習モデルに前記ステップデータを与えて学習結果として前記自移動体と他移動体の前記位置関係を示す状態量と報酬と状態を含むエピソードデータを出力させ、当該エピソードデータをエピソードデータ蓄積部に蓄積する強化学習部と、

前記エピソードデータ蓄積部に蓄積された前記エピソードデータを分析して学習パターンを生成するエピソード分析部と、

前記学習パターンに該当する前記エピソードデータを前記エピソードデータ蓄積部から取得して学習用エピソードデータを生成し、前記機械学習モデルに前記学習用エピソードデータを与えて学習させるエピソードデータ学習部と、

を有し、

前記強化学習部は、

前記ステップデータの開始から所定の終了条件に達するまでのステップデータを１つのエピソードとし、前記エピソードの前記ステップデータに対応するステップを前記エピソードデータに設定し、前記エピソードデータの前記状態には前記ステップデータ毎のシミュレーション結果を設定し、前記エピソードには前記シミュレーション結果に応じて所定の目的を達成したか否かを示すエピソード種別を設定し、
前記エピソード分析部は、
前記エピソード種別と前記シミュレーション結果と前記エピソードの数と前記エピソードデータのステップ数に基づいて前記学習パターンを生成し、
前記エピソード分析部は、
前記シミュレーション結果毎に前記エピソードの数の範囲と、前記エピソードデータのステップ数の範囲を設定し、
前記エピソードデータ学習部は、
前記エピソードデータ蓄積部から、前記シミュレーション結果毎に前記エピソードの数の範囲と、前記エピソードデータのステップ数の範囲に該当するエピソードを取得して前記機械学習モデルに学習させ、
前記エピソードデータ学習部は、
前記シミュレーション結果毎に、前記エピソードの数とステップ数の範囲で抽出したエピソードを、前記エピソードの順序で所定の置換を行う、ことを特徴とする行動制御計画装置。

【請求項２】
請求項１に記載の行動制御計画装置であって、
前記エピソードデータ学習部は、
前記取得したエピソードに対応する前記エピソードデータを前記機械学習モデルに直接入力して学習させることを特徴とする行動制御計画装置。

【請求項３】
請求項１に記載の行動制御計画装置であって、
前記所定の目的は、
前記自移動体が前記他移動体の追跡であることを特徴とする行動制御計画装置。

【請求項４】
請求項１に記載の行動制御計画装置であって、
前記エピソード分析部は、
前記エピソードの数の範囲と、前記エピソードデータのステップ数の範囲において、前記シミュレーション結果の種別に応じた所定の比率で前記エピソードを割り当てることを特徴とする行動制御計画装置。

【請求項５】
プロセッサとメモリを有する計算機が、他移動体の行動に対して自移動体の行動を計画する行動制御計画方法であって、
前記計算機が、前記自移動体と他移動体の行動をシミュレーション条件として設定するシミュレーション条件設定ステップと、
前記計算機が、前記シミュレーション条件と予め設定された機械学習モデルに基づいて、前記自移動体と他移動体の位置関係を所定の時間間隔毎にステップデータとして出力するシミュレーションステップと、
前記計算機が、前記ステップデータを取得して、前記機械学習モデルに前記ステップデータを与えて学習結果として前記自移動体と他移動体の前記位置関係を示す状態量と報酬と状態を含むエピソードデータを出力させ、当該エピソードデータをエピソードデータ蓄積部に蓄積する強化学習ステップと、
前記計算機が、前記エピソードデータ蓄積部に蓄積された前記エピソードデータを分析して学習パターンを生成するエピソード分析ステップと、
前記計算機が、前記学習パターンに該当する前記エピソードデータを前記エピソードデータ蓄積部から取得して学習用エピソードデータを生成し、前記機械学習モデルに前記学

習用エピソードデータを与えて学習させるエピソードデータ学習ステップと、
を含み、
前記強化学習ステップは、
前記ステップデータの開始から所定の終了条件に達するまでのステップデータを１つのエピソードとし、前記エピソードの前記ステップデータに対応するステップを前記エピソードデータに設定し、前記エピソードデータの前記状態には前記ステップデータ毎のシミュレーション結果を設定し、前記エピソードには前記シミュレーション結果に応じて所定の目的を達成したか否かを示すエピソード種別を設定し、
前記エピソード分析ステップは、
前記エピソード種別と前記シミュレーション結果と前記エピソードの数と前記エピソードデータのステップ数に基づいて前記学習パターンを生成し
前記エピソード分析ステップは、
前記シミュレーション結果毎に前記エピソードの数の範囲と、前記エピソードデータのステップ数の範囲を設定し、
前記エピソードデータ学習ステップは、
前記エピソードデータ蓄積部から、前記シミュレーション結果毎に前記エピソードの数の範囲と、前記エピソードデータのステップ数の範囲に該当するエピソードを取得して前記機械学習モデルに学習させ、
前記エピソードデータ学習ステップは、
前記シミュレーション結果毎に、前記エピソードの数とステップ数の範囲で抽出したエピソードを、前記エピソードの順序で所定の置換を行う、ことを特徴とする行動制御計画方法。
【請求項６】
プロセッサとメモリを有する行動制御計画装置と、
前記行動制御計画装置に接続されて所定のシミュレーションを行うシミュレータと、を有して、他移動体の行動に対して自移動体の行動を計画する行動制御計画システムであって、
前記シミュレータは、
前記行動制御計画装置から受け付けたシミュレーション条件で前記シミュレーションを実施し、
前記行動制御計画装置は、
前記自移動体と他移動体の行動をシミュレーション条件として設定し、前記シミュレーション条件と予め設定された機械学習モデルを前記シミュレータに送信して、前記自移動体と他移動体の位置関係を所定の時間間隔毎にステップデータとして出力させるシミュレーション条件設定部と、
前記ステップデータを取得して、前記機械学習モデルに前記ステップデータを与えて学習結果として前記自移動体と他移動体の前記位置関係を示す状態量と報酬と状態を含むエピソードデータを出力させ、当該エピソードデータをエピソードデータ蓄積部に蓄積する強化学習部と、
前記エピソードデータ蓄積部に蓄積された前記エピソードデータを分析して学習パターンを生成するエピソード分析部と、
前記学習パターンに該当する前記エピソードデータを前記エピソードデータ蓄積部から取得して学習用エピソードデータを生成し、前記機械学習モデルに前記学習用エピソードデータを与えて学習させるエピソードデータ学習部と、
を有し、
前記強化学習部は、
前記ステップデータの開始から所定の終了条件に達するまでのステップデータを１つのエピソードとし、前記エピソードの前記ステップデータに対応するステップを前記エピソードデータに設定し、前記エピソードデータの前記状態には前記ステップデータ毎のシミュレーション結果を設定し、前記エピソードには前記シミュレーション結果に応じて所定

の目的を達成したか否かを示すエピソード種別を設定し、
前記エピソード分析部は、
前記エピソード種別と前記シミュレーション結果と前記エピソードの数と前記エピソードデータのステップ数に基づいて前記学習パターンを生成し、
前記エピソード分析部は、
前記シミュレーション結果毎に前記エピソードの数の範囲と、前記エピソードデータのステップ数の範囲を設定し、
前記エピソードデータ学習部は、
前記エピソードデータ蓄積部から、前記シミュレーション結果毎に前記エピソードの数の範囲と、前記エピソードデータのステップ数の範囲に該当するエピソードを取得して前記機械学習モデルに学習させ、
前記エピソードデータ学習部は、
前記シミュレーション結果毎に、前記エピソードの数とステップ数の範囲で抽出したエピソードを、前記エピソードの順序で所定の置換を行う、ことを特徴とする行動制御計画システム。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、艦艇、車両といった移動体の行動において、特に他の移動体との位置関係を鑑みて行動する追尾や哨戒といった行動において、他移動体に対する自移動体の行動制御方法を自動的に計画する行動制御計画装置に関する。
【背景技術】
【０００２】
環境に応じて移動体の移動方法、すなわち行動制御内容を決めるためには、様々な目的、要件を鑑みることが必要である。環境としては、移動範囲の制限、障害物といった固定的な制限の他にも、他の移動体との位置関係による制限もある。特に、他の移動体との位置関係については、それが障害物であって回避すればよいといった目的の他にも、他の移動体との距離を一定に保つ、もしくは追跡、追尾する、また他の移動体を探索、検知するといった目的もある。
【０００３】
制御対象である自移動体（ｏｗｎ　ｖｅｈｉｃｌｅ）の行動制御内容を決めるためには、従来は制御理論などに基づき制御方法が設計されるが、強化学習を利用することで行動制御内容を計画できる。近年、状態量と行動（方策）、行動価値もしくは報酬の関係で定義される、深層ニューラルネットワークを訓練することで強化学習を行う、深層強化学習といった方法により学習能力が向上しており、強化学習技術の適用が検討される。
【０００４】
特許文献１では、障害物が設定された環境において移動物の行動を決定するために、深層強化学習により、移動物の行動を決定する技術が提案される。
【０００５】
特許文献２では、特に深層強化学習技術に関して、深層ニューラルネットワークを訓練する際、ステップ毎の状態量、行動、報酬のデータに対し、学習の効果が高くなる指標である期待学習進展度を計算して関連付け、リプレイメモリ（訓練のためのステップデータの蓄積）から優先順位付けして訓練に用いる技術が提案される。
【先行技術文献】
【特許文献】
【０００６】
【文献】特開２０１８－１９８０１２号公報
【文献】特開２０２０－４７２８５号公報
【発明の概要】
【発明が解決しようとする課題】

【０００７】
環境に応じて移動体の行動制御内容を強化学習技術を利用して決定する場合、学習を実現することが問題となる。移動体の行動の目的を達成するための前提とする環境などの条件が単純であれば、学習は容易であるが、条件が複雑であれば、学習にかかる訓練の回数が多く必要となり、さらに学習自体を達成できなくなる。
【０００８】
特に移動体の行動制御計画では、追跡、探索といった目的もあり、他の移動体（以下、他移動体）との位置関係を反映することが必要となる。この場合、自移動体の位置、移動方向に対して、近づく、離れる、横切るなど、他移動体の位置、移動方向が多様であるが、いずれに対しても適切な行動をとれることが重要である。
【０００９】
特許文献１では、環境は動かない障害物、もしくは通り道といった設定であり、移動体の一つの学習結果から他の多様な環境に対応する学習方法が示されない。
【００１０】
特許文献２では、深層ニューラルネットワークの訓練を効率化して、学習を成功させるための技術である。環境に応じた学習をするためには、成功か失敗に至った時点までを示すエピソード自体の設定を変えて学習しなければならないが、その設定方法に関する工夫は示されない。
【００１１】
本発明の目的は、多様な環境に対して、特に、自移動体に対して他移動体が様々な行動を行うことに対応して、自移動体が行動できるように行動を計画することである。この目的のため、他移動体の異なる行動に対して強化学習を行った結果のエピソードのデータを利用して、エピソードデータを用いて自移動体の行動計画を学習する技術を提供するものである。
【課題を解決するための手段】
【００１２】
本発明は、プロセッサとメモリを有して、他移動体の行動に対して自移動体の行動を計画する行動制御計画装置であって、前記自移動体と他移動体の行動をシミュレーション条件として設定するシミュレーション条件設定部と、前記シミュレーション条件と予め設定された機械学習モデルに基づいて、前記自移動体と他移動体の位置関係を所定の時間間隔毎にステップデータとして出力するシミュレータと、前記ステップデータを取得して、前記機械学習モデルに前記ステップデータを与えて学習結果として前記自移動体と他移動体の前記位置関係を示す状態量と報酬と状態を含むエピソードデータを出力させ、当該エピソードデータをエピソードデータ蓄積部に蓄積する強化学習部と、前記エピソードデータ蓄積部に蓄積された前記エピソードデータを分析して学習パターンを生成するエピソード分析部と、前記学習パターンに該当する前記エピソードデータを前記エピソードデータ蓄積部から取得して学習用エピソードデータを生成し、前記機械学習モデルに前記学習用エピソードデータを与えて学習させるエピソードデータ学習部と、を有する。
【発明の効果】
【００１３】
本発明によると、様々な環境に対して、特に自移動体に対して他移動体の行動が多様な関係となる場合でも、例えば追跡、検知といった目的を達成するための、自移動体の行動を学習できる。
【００１４】
本明細書において開示される主題の、少なくとも一つの実施の詳細は、添付されている図面と以下の記述の中で述べられる。開示される主題のその他の特徴、態様、効果は、以下の開示、図面、請求項により明らかにされる。
【図面の簡単な説明】
【００１５】
【図１】本発明の実施例を示し、行動制御計画装置で行われる処理の一例を示した図であ

る。
【図２】本発明の実施例を示し、強化学習と、エピソード分析及びエピソードデータ学習による行動制御計画の処理の一例を示した図である。
【図３】本発明の実施例を示し、移動体の行動制御計画のフローチャートの一例を示した図である。
【図４】本発明の実施例を示し、移動体の行動制御計画を実行する行動制御計画装置の構成の一例を示した図である。
【図５】本発明の実施例を示し、追跡における自移動体、他移動体の行動環境の例を示した図である。
【図６】本発明の実施例を示し、哨戒における自移動体、他移動体の行動環境の例を示した図である。
【図７Ａ】本発明の実施例を示し、哨戒領域の構成の例を示した図である。
【図７Ｂ】本発明の実施例を示し、他移動体の行動パターンの例を示した図である。
【図８】本発明の実施例を示し、方策関数、状態価値関数の深層ニューラルネットワークによる構成の例を示した図である。
【図９Ａ】本発明の実施例を示し、エピソードデータの構成の失敗例を示した図である。
【図９Ｂ】本発明の実施例を示し、エピソードデータの構成の成功例を示した図である。
【図１０】本発明の実施例を示し、一連の強化学習の手順の、フローチャートの一例を示した図である。
【図１１Ａ】本発明の実施例を示し、１つの強化学習過程についてのエピソードデータの例を示した図である。
【図１１Ｂ】本発明の実施例を示し、１つの強化学習過程についてのエピソードデータの例を示した図である。
【図１２】本発明の実施例を示し、多くの成功した強化学習の学習過程についての、ステップ数のプロットの例を示した図である。
【図１３】本発明の実施例を示し、強化学習過程についての、失敗、成功エピソード別のステップ数の関係の例を示した図である。
【図１４】本発明の実施例を示し強化学習過程における失敗、成功エピソードの出現の範囲の一例を示した図である。
【図１５】本発明の実施例を示し、学習パターンの一例を示した図である。
【図１６】本発明の実施例を示し、強化学習過程に対するステップ数のプロットとして、学習パターンのグラフの一例を示した図である。
【図１７】本発明の実施例を示し、学習パターンに対応するエピソードデータのサンプリング結果の一例を示した図である。
【図１８Ａ】本発明の実施例を示し、エピソードデータのサンプリング結果を置換した例を示した図である。
【図１８Ｂ】本発明の実施例を示し、エピソードデータのサンプリング結果を置換した他の例を示した図である。
【図１９】本発明の実施例を示し、一連のエピソードデータ学習の処理手順の、フローチャートの一例を示した図である。
【発明を実施するための形態】
【００１６】
以下、本発明の実施形態を添付図面に基づいて説明する。
【００１７】
自律的に移動する移動体（自移動体）の行動制御計画のためには、自移動体と他移動体を移動させるシミュレーションによる強化学習技術を利用する。強化学習では、自移動体と他移動体の位置関係の初期値から、特定の目的の達成について成功、もしくは失敗するまで１ステップずつ行動する。なお、ステップは所定の時間間隔毎に実行するシミュレーション（又は学習や認識）の処理を示す。
【００１８】

自移動体と他移動体を移動させるシミュレーションが開始されてから、自移動体の行動が成功又は失敗に至った時点までを１つのエピソードと言う。行動制御計画装置は、エピソードを繰り返すことで目的を達成する行動制御を機械学習モデルで学習する。強化学習のためには状態量と行動（方策）、状態価値の関係は深層ニューラルネットワーク（機械学習モデル）で定義され、行動制御計画装置では、エピソード毎に状態、行動、報酬のステップ毎のデータを用いて深層ニューラルネットワークが強化学習とエピソードデータ学習によって訓練される。

【００１９】

なお、エピソードを構成するステップは、例えば、所定時間間隔毎の状態量と、行動と、状態価値が含まれる。そして、状態量には、自移動体や他移動体の位置や速度を含むことができ、行動には自移動体の進路や速度を含むことができ、状態価値には、ステップ毎の報酬を含むことができる。なお、他移動体は１つ以上の移動体で、複数の移動体も対象とすることができる。

【００２０】

状態量と行動（方策）、状態価値との関係は損失関数を最小化する最適化により訓練されるので、最適化を効率化するステップデータを優先付けて、選別する。またステップの時系列の並びであるエピソードは、学習完了までの順序の並びとなる。なお、損失関数は、深層ニューラルネットワークの予測精度を評価する関数であり、例えば、回帰モデル等の周知又は公知の技術を適用することができる。

【００２１】

自移動体に対する他移動体の行動パターンが単純、もしくは一通りである場合、強化学習を適用すれば学習は達成される。しかし他移動体の行動パターンは様々であることが想定され、一通りであることを仮定できない。

【００２２】

そこで行動制御計画装置は、他移動体と自移動体の多様な行動パターンについて、各行動パターン別に強化学習を行ったときのエピソードデータを蓄積する。エピソードデータとは、ステップ毎の状態量、行動（方策）、状態価値の並びで構成されたデータである。また、エピソードデータは、複数のステップデータを含むことができる。本実施例では予め蓄積したエピソードデータより、学習が成功に至るまでのエピソード毎の学習過程を分析する。

【００２３】

行動制御計画装置は、分析の結果として、ステップ数に応じた失敗のエピソードと成功のエピソードの、学習開始からのエピソードの順番である、学習パターンを設定する。なお、行動制御計画装置は、１つのエピソードに含まれるステップ数の多寡に応じて学習パターンを設定することができる。

【００２４】

状態量を入力とし、行動（方策）、状態価値（報酬）を出力として、それらの関係が定義される深層ニューラルネットワークは、予め蓄積されたエピソードデータを用いて訓練される。予め蓄積されたエピソードデータを用いてもステップ毎、エピソード毎に深層ニューラルネットワークを訓練できる。この訓練の際には強化学習で利用されるランダムな行動選択はしない。このエピソードデータを用いた深層ニューラルネットワークの訓練による学習をエピソードデータ学習と呼ぶ。

【００２５】

エピソードデータ学習では学習パターンの設定に応じて、効率的に深層ニューラルネットワークを訓練でき、例えば、少ないエピソード数で学習が完了する、といった観点で効率化でき、学習結果の深層ニューラルネットワーク１０によるシミュレーション結果は成功しやすくなる。

【００２６】

行動制御計画装置は、個別のシミュレーション結果の強化学習の結果で得られたエピソードデータを利用し、複数のシミュレーション結果（例えば、拿捕、見失い、好位置など

）について予め学習パターンを設定して、学習パターンに一致するエピソードデータを予め蓄積されたエピソードデータから抽出してエピソードデータ学習を行う。

【００２７】

なお、他移動体のシミュレーション結果は、上記の３種類に限定されるものではなく、予め設定された種類のシミュレーション結果を設定すればよい。シミュレーション結果は、自移動体と他移動体の位置関係を初期値から変化させて、所定の終了条件を満足した時点（エピソードの終了時点）の他移動体の位置から後述するように判定することができる。

【００２８】

これにより、行動制御計画装置は、他移動体の複数の行動パターンに対して、状態量に対する行動（方策）の選択を深層ニューラルネットワークで学習する。すなわち、深層ニューラルネットワークは他移動体の行動パターンに応じた自移動体の行動制御計画を学習する。学習の完了は、複数のシミュレーション結果の全てに対して、エピソードの処理を試行し、各試行が成功することをもって判断する。

【００２９】

以上により、シミュレーション結果毎の強化学習の結果であるエピソードデータを利用して、強化学習が成功するに至るエピソード毎の成功と失敗のステップ数を分析して学習パターンを設定する。

【００３０】

そして、行動制御計画装置は、複数のシミュレーション結果についてエピソードデータ学習をすることにより、複数の行動パターンに対する自移動体の行動制御方法を計画できる。本実施例の行動制御計画装置は、従来の強化学習だけでは実現することが困難な、自移動体と他移動体の行動の複雑な関係において、深層ニューラルネットワークに効率よく行動制御を学習させることが可能となる。

【００３１】

＜処理の概要＞

図１の行動制御計画装置の処理の一例を示す図により、本発明の一実施の形態に係わる、強化学習、エピソード分析、エピソードデータ学習による、行動制御計画の処理の概要を示す。まず、行動制御計画装置１０５で行われる処理の全体の流れを図２のフローチャートにより説明する。

【００３２】

多様な環境や、他移動体の行動パターンに対して、自移動体の行動制御を計画するために、行動制御計画装置１０５は、まず、個別のシミュレーションの条件に対して自移動体の行動制御の強化学習を行う（２０１）。この強化学習においては、各移動体の状態に応じた深層ニューラルネットワーク１０による行動選択を１ステップずつ進めて、自移動体の目的が達成されて成功、もしくは失敗するまでの１つのエピソードの処理において、状態量と方策、状態価値の関係を定める関数である深層ニューラルネットワーク１０の訓練を行う。

【００３３】

行動制御計画装置１０５は、エピソードの処理を繰り返して、必ず、もしくは十分に目的が達成される関数を得ることで学習を完了する。行動制御計画装置１０５は、学習が完了に至るまでのエピソードのステップ毎の状態量と、行動及び、報酬のデータを、強化学習過程の順序に対する全てのエピソードによる学習結果として、エピソードデータベース１２２に蓄積する（２０２）。

【００３４】

行動制御計画装置１０５（図１参照）は様々な行動パターンに対する学習結果を蓄積する。また同一の学習パターンに対して、一つの学習結果だけでなく、多数の学習結果を蓄積する。これらの学習結果の一つのエピソードのデータをエピソードデータと呼ぶ。

【００３５】

行動制御計画装置１０５は、エピソードデータベース１２２に予め蓄積されたエピソードデータに対して、個別の条件毎に強化学習を実施してエピソードデータを収集し、学習

過程の順にエピソードの種別（成功と失敗）、エピソードのステップ数を集計して、学習過程のエピソードを分析する（２０３）。

【００３６】

これにより、行動制御計画装置１０５は、学習過程のエピソード順の、成功と失敗と、それらのステップ数を集計して学習パターンを設定する。

【００３７】

そして、行動制御計画装置１０５は、複数の条件についての学習パターンの設定に従って、エピソードデータをエピソードデータベース１２２から収集する。そして、行動制御計画装置１０５は、学習過程の順にエピソードデータを用いて状態量と方策、状態価値の関数（深層ニューラルネットワーク）の訓練を行う、エピソードデータ学習を行う（２０４）。学習過程の順のエピソード完了毎に、移動体の目的の達成（シミュレーション結果）を判定することで学習の完了を判定する。

【００３８】

行動制御計画装置１０５は、エピソードデータ学習の学習結果を蓄積する（２０５）。エピソードデータ学習におけるエピソードデータは、既にエピソードデータベース１２２に蓄積されたものであり、エピソードデータそのものの蓄積は不要であり、学習過程で利用したエピソードデータのリストと、エピソードデータ学習における、エピソード毎の移動体の目的達成の判定結果（エピソードの種別）が蓄積されればよい。

【００３９】

以上が、図２のフローチャートによる、強化学習、エピソード分析、エピソードデータ学習による、行動制御計画の処理の全体の流れの説明である。上記処理によって、強化学習で出力されたエピソードデータがエピソードデータベース１２２に蓄積され、エピソード分析で設定された学習パターンに応じたエピソードデータによって深層ニューラルネットワーク１０が訓練され、訓練の結果はエピソードデータ学習で利用したエピソードデータのリストと、エピソード毎の移動体の目的達成の判定結果が蓄積されていく。

【００４０】

図１の処理の一例を示す図は、本実施例の処理の概要において、必要となる処理機能と、処理の順序と、データの授受を矢印で示したものである。図中縦方向を強化学習部１０１、エピソード分析部１０２、エピソードデータ学習部１０３で分類し、図中横方向を処理の主体として、オペレータ１０４と、行動制御計画装置１０５で分類している。オペレータ１０４の処理とは利用者などのマニュアルによる処理を意味する。

【００４１】

強化学習部１０１においては、自移動体が移動する空間の環境や、他移動体の行動パターン、また自移動体の移動方向、速度といった行動の設定などのシミュレーションを行う条件設定（シミュレーション条件設定部）１１１を処理する。強化学習部１０１は、オペレータ１０４から指定された条件（シミュレーション条件）で、シミュレータ１１２により他移動体と、自移動体の行動をシミュレーションする。

なお、シミュレータ１１２は、既に学習が完了した深層ニューラルネットワーク（機械学習モデル）を用いてシミュレーションを実施する。なお、シミュレータ１１２は、周知又は公知の技術を適用すればよく、移動体の種類に応じたシミュレータ１１２を使用することができる。また、移動体の種類は同一である必要はなく、例えば、自移動体が艦船、他移動体が潜水艦でもよいし、自移動体が航空機、他移動体が艦船であってもよい。

【００４２】

シミュレータ１１２では、ステップ（所定の時間間隔）毎に状態更新１１３と行動選択１１４の処理がなされる。特に行動選択１１４では、強化学習のために、状態量と方策の関数から行動を選択するだけでなく、行動を乱択する処理も行う。

【００４３】

自移動体の目標の成功、失敗までの一つのエピソードはステップ毎の処理を含み、ステップの状態量、行動、報酬のデータを用いて状態量と方策、行動価値の関数である深層ニューラルネットワーク１０の訓練を、学習器１１５で処理する。

【００４４】
シミュレーション結果のステップ毎の状態量、行動、報酬のデータは行動制御計画装置１０５のメモリ（記憶装置）に蓄積されており、訓練データ取得機能１１６により訓練用のデータがシミュレータ１１２から取得され、深層ニューラルネットワーク訓練１１７を処理する。自移動体の目的が達成されるように深層ニューラルネットワーク１０の訓練が進み、強化学習が完了したら、強化学習部１０１における処理過程についてのエピソードデータが、エピソードデータベース（エピソードデータ蓄積部）１２２に蓄積される。
【００４５】
エピソード分析部１０２においては、エピソードデータベース１２２から、エピソードの種別や他移動体のシミュレーション結果毎に、強化学習過程のエピソードデータを収集して、エピソード分析機能１２１により、学習パターンを生成する。
【００４６】
エピソードデータ学習部１０３では、対象とする複数の他移動体のシミュレーション結果について生成された学習パターンを用いて、エピソードデータ学習用に学習パターン設定１３１の処理を行う。
【００４７】
そして、学習パターン設定１３１で設定された学習パターンによりエピソード選択１３２の処理を行って、エピソードデータ学習１３３の処理を行う。エピソードデータ学習においても深層ニューラルネットワーク１０の訓練のためには、学習器１１５を用いて処理を行う。
【００４８】
なお、条件設定１１１と学習パターン設定１３１はオペレータ１０４によるマニュアル設定を想定して図１に示したが、設定される条件を規則的に定めたり、学習パターンの設定方法が予め定められる場合には、行動制御計画装置１０５の処理としてもよい。
【００４９】
以上が図１の行動制御計画装置１０５で行われる処理の説明である。
【００５０】
＜行動制御計画＞
本発明の、移動体の行動制御計画の一例を示す。
【００５１】
図３のフローチャートにより、移動体の行動制御計画方法の一例を示す。フローチャートの処理は、強化学習部１０１、エピソード分析部１０２、エピソードデータ学習部１０３に分割して示している。各処理は行動制御計画装置１０５が実行するものとして記載する。
【００５２】
強化学習部１０１においては、まず、環境、他移動体の行動パターン、自移動体の行動等のシミュレーション条件を受け付けて、シミュレーション及び強化学習のためのパラメータについてのシミュレーション条件を取得する（３０１）。シミュレーション条件は、図１で示したようにオペレータ１０４が設定することができる。
【００５３】
状態量に対する方策と、行動価値の関数である深層ニューラルネットワーク１０を訓練する処理では、強化学習のためのエピソードの繰り返しのステップ３０２からステップ３０４の範囲を繰り返して処理する。繰り返しの一回は一つのエピソードの処理のこととなる。
【００５４】
行動制御計画装置１０５は、繰り返しの１回について、１エピソードの各ステップについて行動選択及び状態変化を深層ニューラルネットワーク１０に学習させる（３０３）。１つのエピソードの処理は、移動体の初期状態から、移動体の目的が達成されるか否かの成功、失敗までの、ステップ毎の処理である。
【００５５】

行動制御計画装置１０５は、ステップ毎、またエピソードの終了時点で、各ステップにおける状態量、行動、報酬のデータを用いて、深層ニューラルネットワーク１０を訓練する。訓練時の行動選択では、行動は深層ニューラルネットワーク１０による行動選択だけでなく、乱択（ランダムによる行動選択）も行う。

【００５６】

また、行動制御計画装置１０５は、１つのエピソードの処理の終了時点で、訓練された深層ニューラルネットワーク１０を用いた行動選択によるエピソードの処理をテスト（試行）し、所定の目的を達成したらエピソードの種別を成功として学習完了とする。この場合、例えば、エピソードが５回連続でテストに成功したら学習完了とする、といった学習完了の判定でもよい。

【００５７】

行動制御計画装置１０５は、深層ニューラルネットワークの学習が完了したら、学習結果の深層ニューラルネットワークを用いてエピソードの試行（３０５）を行うことにより、学習が完了したこと、また移動体の行動内容を確認する。もしくは、学習が完了したことを確認できるようにログなどの記録を残すことができる。

【００５８】

そして強化学習部１０１は、強化学習の結果（エピソード）のデータベース蓄積（３０６）で、強化学習での学習過程の全てのエピソードの、ステップ毎の状態量、行動、報酬のデータを、学習過程についてのエピソードデータとしてエピソードデータベース１２２に蓄積する。なお、強化学習部１０１は、エピソードデータを蓄積する際には、他移動体のシミュレーション結果とエピソードの種別を含めてエピソードデータベース１２２へ蓄積する。

【００５９】

なお、このフローチャートにおいて、強化学習部１０１は、一通りのフローで示しているが、他移動体の行動パターンについての強化学習を、同一の行動パターンも含み、多数回実行する。

【００６０】

エピソード分析部１０２においては、まず他移動体のシミュレーション結果毎に、エピソードデータベース１２２からエピソード収集（分析用）（３０７）を行う。この際には、強化学習における学習過程の順序についての情報も取得する。

【００６１】

学習成功のエピソード失敗／成功過程収集（３０８）では、強化学習を実施した深層ニューラルネットワーク１０がシミュレーションで成功した学習過程について、エピソード分析部１０２がエピソードの成功と失敗、そしてエピソードデータのステップ数を収集し、一覧化する。そして、学習過程におけるエピソードデータの順に、エピソード順の間隔毎に、失敗と成功の件数と各エピソードデータのステップ数を集計する。

【００６２】

行動制御計画装置１０５は、エピソードデータの集計結果より、エピソード順の成功と失敗のサンプル（エピソードデータ）の件数と、ステップ数の範囲を設定して、学習パターンを構成する（３０９）。

【００６３】

なお、このフローチャートにおいてエピソード分析部１０２は、一通りのフローで示しているが、他移動体の異なる行動パターンについても同様にエピソード分析を行う。

【００６４】

エピソードデータ学習部１０３においては、まずステップ３０１と同様の条件と併せて、学習パターンを含めて、条件（エピソード条件）を取得（学習パターンを含む）する（３１０）。なお、この条件（エピソード条件）には、複数の他移動体のシミュレーション結果の指定も含むことができる。

【００６５】

そしてエピソード学習のために、学習パターンで指定されている学習過程の順序におけ

るエピソードの種別（成功と失敗）、及びステップ数に該当するエピソードデータをエピソードデータベース１２２から検索して取得する、エピソード収集（学習用）を行う（３１１）。

【００６６】

エピソードデータ学習部１０３において、状態量に対する方策と、行動価値の関数である深層ニューラルネットワークを訓練する処理は、収集したエピソードについて、ステップ３１２からステップ３１４の範囲でエピソード毎の繰り返しで行う。エピソード毎の繰り返しは、予め設定された終了条件（例えば、処理済みのエピソード数等）が成立するまで行うことができる。

【００６７】

行動制御計画装置１０５は、繰り返しの１回について、１エピソードの行動選択及び状態データを学習する（３１３）。ここでの学習は、行動を乱択することもある、いわゆる強化学習ではなく、エピソードデータを直接に用いた学習である。１つのエピソードデータは、移動体の初期状態から、エピソードが成功又は失敗するまでの、ステップ毎の処理である。

【００６８】

エピソードデータ学習部１０３は、エピソードデータのステップ毎に処理を進めるが、シミュレータ１１２を利用して移動体の行動を選択したり、状態を更新して、状態量、報酬を計算することはない。

【００６９】

エピソードデータ学習部１０３は、ステップ毎にエピソードデータの状態量、行動、報酬を各移動体に設定していく。このデータをメモリに格納し、ステップ毎、またエピソードの終了時点で、状態量、行動、報酬のデータを用いて、深層ニューラルネットワークを訓練する。

【００７０】

エピソードデータ学習部１０３は１つのエピソードの処理の終了時点で、訓練された深層ニューラルネットワークを用いた行動選択によるエピソードの処理をテストする。この際に、全ての他移動体のシミュレーション結果に対するエピソードの処理をテストし、全て成功したら学習完了とする。また各シミュレーション結果についてエピソードが５回連続で成功したら学習完了とする、といった終了条件による判定をしてもよい。

【００７１】

学習が完了したら、エピソードデータ学習部１０３は、学習結果でのエピソード試行（３１５）を行うことにより、学習が完了したこと、また各移動体の行動内容を確認する。もしくは確認できるようにログなどの記録を出力する。エピソードデータ学習部１０３は、学習で指定された、他移動体のシミュレーション結果の全てに対して試行する。

【００７２】

エピソードデータ学習部１０３は、エピソードデータ学習結果のデータベース蓄積（３１６）を行う。なお、学習に利用したエピソードデータの登録はしない。エピソードデータ学習の学習過程、例えば学習完了までにかかったエピソード数や利用したエピソードデータなどをエピソードデータベース１２２に蓄積する。

【００７３】

以上が、図３のフローチャートによる、移動体の行動制御計画方法の一例の説明である。上記処理によって、深層ニューラルネットワーク１０が出力したエピソードデータが強化学習の結果としてエピソードデータベース１２２に格納され、エピソード分析部１０２は、エピソードの成功と失敗及びエピソードデータのステップ数を収集し、エピソードの成功と失敗の件数と、ステップ数の範囲を設定して学習パターンを後述するように生成する。

【００７４】

エピソードデータ学習部１０３は、学習パターンに対応するエピソードデータをエピソードデータベース１２３から選択して学習器１１５を用いてエピソードデータ学習１３３

を実施して深層ニューラルネットワーク１０訓練を実施する。そして、エピソードデータ学習１３３は、訓練された深層ニューラルネットワーク１０について試行を実施してからエピソードデータベース１２２にエピソード数等の学習過程のデータを蓄積する。

【００７５】

＜行動制御計画装置の構成＞

次に、図４の行動制御計画装置１０５の機能の一例を説明する。図４は、図３のフローチャートで説明した内容を実現する行動制御計画装置１０５の機能である。

【００７６】

行動制御計画装置１０５は、プロセッサ１１と、メモリ１２と、ストレージ装置１３と、入出力装置１４を含む計算機で構成される。なお、入出力装置１４は、キーボードやマウス又はタッチパネル等の入力装置と、ディスプレイなどの出力装置を含む。

【００７７】

行動制御計画装置１０５のメモリ１２には、シミュレータ４００と、強化学習／エピソードデータ学習部４１０と、エピソード分析部４３０の機能部がプログラムとしてロードされて、プロセッサ１１によって実行される。なお、各プログラムが利用するデータ（行動環境データベース４１８、エピソードデータベース４１９等）はストレージ装置１３に格納してもよい。

【００７８】

強化学習／エピソードデータ学習部４１０は、深層ニューラルネットワーク（ＤＮＮ：Ｄｅｅｐ　Ｎｅｕｒａｌ　Ｎｅｔｗｏｒｋ）１０の訓練に関する処理が、強化学習部（図１の１０１に相当）とエピソードデータ学習部（図１の１０３に相当）と共通であるため、強化学習部とエピソードデータ学習部を合わせて一つの機能ブロック（４１０）とした。しかし、エピソードデータ学習部は強化学習とは異なる学習方法であるため、機能ブロック（４１０）の内側に、エピソードデータ学習部４２０の機能ブロックを配置した。

【００７９】

シミュレータ４００（図１の１１２に相当）の機能は強化学習／エピソードデータ学習部４１０で利用される。シミュレータ４００は、図３における１エピソードの行動選択・状態変化の学習（３０３）の処理において、ステップ毎の処理を進めるための機能である。

【００８０】

シミュレータ４００は行動環境設定機能４０１、移動体行動処理機能４０２、移動体状態量算出機能４０３、移動体状態判定機能４０４を有する。

【００８１】

行動環境設定機能４０１は、自移動体、他移動体の初期位置や障害物などの環境を設定する機能であり、本機能により初期状態を決める。

【００８２】

移動体行動処理機能４０２により、ステップ毎で自移動体の行動を選択し、移動処理を行う。移動体行動処理機能４０２他移動体の移動も行い、これによりステップ毎の状態が更新されることとなる。移動体行動処理機能４０２は、図１の状態更新１１３と、行動選択１１４に対応する。

【００８３】

移動体状態量算出機能４０３により、ステップ毎で状態が更新された際に、状態量を計算する。状態量とは、自移動体が予め設定された目的を達成するために必要となる情報であって、例えば、他移動体との距離や、障害物との距離、他移動体の方位、他移動体の速度、などである。

【００８４】

移動体状態判定機能４０４により、計算された状態量から自移動体のそのステップにおける状態を判定する。状態は、例えば自移動体と他移動体の距離によれば、他移動体を検知できている良好な位置にある状態（好位置）、他移動体を検知できない距離となって見失った状態、検知できているが近くなりすぎて他移動体に見つかり拿捕される状態、といった記号的な状態名称の判定が行われる。なお、本実施例の自移動体は、艦船の例を示す。

【００８５】
移動体状態量算出機能４０３は、この判定結果を利用して、他移動体の見失い、自移動体の拿捕といった状態ならばエピソードは失敗、また所定ステップ数を満たして良好な位置（好位置）ならば成功と判定しエピソードを終了する。また、報酬の値は、この状態に基づいて設定される。
【００８６】
以上の機能により、図３のステップ３０３の処理（行動選択及び状態変化の学習）が実現される。なおシミュレーションを実行することは強化学習では必要だが、エピソードデータ学習そのものには不要な機能である。しかし、一般にシミュレータ４００は自移動体、他移動体、環境に関するデータを保持し、エピソードデータにアクセスする機能をプログラムとして有しており、当該プログラムの機能はエピソードデータ学習でも利用され得る。
【００８７】
強化学習／エピソードデータ学習部４１０の枠内は学習を実現するための各機能を示している。このうち、特に強化学習に係わる機能について示す。強化学習／エピソードデータ学習部４１０は、方策・状態価値関数構成機能４１１、エピソード処理機能４１２、報酬設定機能４１３、ＤＮＮ訓練機能４１４、ＤＮＮ訓練用データサンプリング機能４１５、方策・状態価値算出機能４１６、エピソードデータ蓄積機能４１７、行動環境データベース４１８、エピソードデータベース４１９の構成要素を有する。
【００８８】
方策・状態価値関数構成機能４１１により、強化学習の開始前に、状態量を入力し、選択される行動、すなわち方策と、状態価値を出力とする深層ニューラルネットワーク１０を構成する。
【００８９】
エピソード処理機能４１２により、１つのエピソードに対して、初期状態からステップ毎に処理を進めて、移動体が目的を達成するか否かによるエピソードの終了までの処理を進める。エピソード処理機能４１２は、図３のステップ３０３の処理を行う機能であり、シミュレータ４００の各機能、及び報酬設定機能４１３、ＤＮＮ訓練機能４１４、ＤＮＮ訓練用データサンプリング機能４１５、方策・状態価値算出機能４１６の呼び出しなど、各種制御を行う機能である。
【００９０】
報酬設定機能４１３は、移動体状態判定機能４０４の結果で得られた状態に応じて報酬を決める機能である。
【００９１】
ＤＮＮ訓練機能４１４により、深層ニューラルネットワーク１０のパラメータを、ステップ毎の状態量、行動、報酬のデータを用いて最適化することで、深層ニューラルネットワーク１０を訓練する。
【００９２】
ＤＮＮ訓練用データサンプリング機能４１５により、深層ニューラルネットワーク１０を訓練するためのデータを、エピソードにおけるステップ毎のデータから特に最適化を効率的に進めるために、サンプリングする。エピソードについては一つのエピソードだけでなく、複数のエピソードに亘るサンプリングとしてもよい。
【００９３】
方策・状態価値算出機能４１６は、深層ニューラルネットワーク１０の入力となる状態量を与えることで、深層ニューラルネットワーク１０の出力である方策と状態価値を取得する機能である。方策を取得するとは、行動を選択するということであり、移動体行動処理機能４０２でこの方策・状態価値算出機能４１６を呼び出すことで行動を決定する。
【００９４】
エピソードデータ蓄積機能４１７により、図３のステップ３０６におけるエピソードデータをエピソードデータベース４１９に蓄積する。

【００９５】
行動環境データベース４１８は、自移動体の行動環境のデータが格納されるデータベースである。
【００９６】
エピソードデータベース４１９は、エピソードデータが格納されるデータベースで、図１のエピソードデータベース１２２に相当する。
【００９７】
エピソードデータ学習部４２０の各機能について説明する前に、図３のフローチャートの順に合わせ、エピソード分析部４３０の各機能について説明する。
【００９８】
エピソード分析部４３０は、エピソードデータ収集機能４３１と、エピソードデータ描画機能４３２、ステップ毎エピソード集計機能４３３、学習パターン構成機能４３４で構成される。なお、エピソード分析部４３０は、図１のエピソード分析部１０２に相当する。
【００９９】
エピソードデータ収集機能４３１により、他移動体のシミュレーション結果毎に、強化学習の学習過程における、エピソードデータの収集を行う。
【０１００】
エピソードデータ描画機能４３２により、エピソードにおける環境、自移動体、他移動体のステップ毎の、例えば配置もしくはマップ（地図）を描画する。またステップ毎の状態量の変化などの描画も含む。マニュアルでのオペレータ１０４によるエピソードの内容確認のための機能である。なお、本実施例では、画像データの生成を描画とする。
【０１０１】
ステップ毎エピソード集計機能４３３により、学習過程における、エピソードの成功と失敗、ステップ数を一覧化し、エピソードの順に成功と失敗の件数とエピソードデータのステップ数を集計する。
【０１０２】
学習パターン構成機能４３４により、エピソード順の間隔毎に、エピソードの種別（成功と失敗）の件数と、エピソードデータのステップ数の範囲を集計することで、学習パターンを構成する。
【０１０３】
強化学習／エピソードデータ学習部４１０でエピソードデータ学習部４２０を構成する各機能について説明する。エピソードデータ学習部４２０は、学習パターン取得機能４２１、エピソードデータ収集・一覧化機能４２２、エピソードデータ学習機能４２３で構成される。
【０１０４】
学習パターン取得機能４２１により、他移動体のシミュレーション結果についての学習パターンを取得する。図３のステップ３１０で利用される機能であり、この際、行動環境設定機能４０１により、各種の条件も取得する。なお、学習パターン取得機能４２１が取得する学習パターンは、エピソード分析部４３０の学習パターン構成機能４３４が出力した学習パターンである。
【０１０５】
エピソードデータ収集・一覧化機能４２２により、学習パターンで指定されたエピソードの条件（エピソード条件）に該当するエピソードデータをエピソードデータベース４１９から収集する。エピソードデータ収集・一覧化機能４２２は収集した結果をエピソードデータ学習で利用されるデータとして一覧化する。
【０１０６】
エピソードデータ学習機能４２３により、エピソードデータ学習の処理を行う。エピソードデータ学習機能４２３は、強化学習におけるエピソード処理機能４１２に相当し、各機能を制御する機能を有する。エピソードデータ学習の開始前には、方策・状態価値関数構成機能４１１により深層ニューラルネットワーク１０を構成する。

【０１０７】
強化学習／エピソードデータ学習部４１０は、エピソードデータのステップ毎のデータをメモリ１２に格納し、ＤＮＮ訓練機能４１４、ＤＮＮ訓練用データサンプリング機能４１５を利用して、深層ニューラルネットワーク１０の訓練を行う。またエピソードデータ学習が完了したら、エピソードデータ蓄積機能４１７により、学習結果をエピソードデータベース４１９に格納する。
【０１０８】
上記各機能はプログラムとしてメモリ１２にロードされてプロセッサ１１によって実行される。プロセッサ１１は、各機能部のプログラムに従って処理を実行することによって、所定の機能を提供する機能部として稼働する。例えば、プロセッサ１１は、行動環境設定プログラムに従って処理を実行することで行動環境設定機能４０１を提供する。他のプログラムについても同様である。さらに、プロセッサ１１は、各プログラムが実行する複数の処理のそれぞれの機能を提供する機能部としても稼働する。計算機及び計算機システムは、これらの機能部を含む装置及びシステムである。
【０１０９】
以上が、図４の構成による、移動体の行動制御計画装置１０５が有する機能の一例の説明である。
【０１１０】
＜移動体の行動＞
本発明における移動体の行動の目的と環境の例を示す。
【０１１１】
自移動体の行動の目的は、他移動体の追跡（追尾）、又は設定された領域への他移動体の侵入を検知する哨戒とする。
【０１１２】
図５に、追跡における自移動体と、他移動体の行動環境の例を示す。
【０１１３】
枠５０１において、２次元空間である平面の座標をとるためのＸ座標軸５０２とＹ座標軸５０３が設定される。なお行動領域は平面を示したが、深さ方向を含む３次元空間としてもよい。
【０１１４】
自移動体５０４は、直進、旋回、また変速して空間を自由に移動できる。初期位置と初期速度は予めシミュレーション条件で設定されるが、方向、速度はステップ毎の行動選択により変化する。
【０１１５】
他移動体５０５は、初期位置から、一定の速度で、一方向５０６に移動する。初期位置、速度、方向は予め設定され、この組み合せで一つの他移動体５０５の行動パターンとなる。
【０１１６】
自移動体５０４には検知距離５０７が設定され、検知距離５０７を半径とする円５０８の内側の他移動体５０５を検知する。追跡開始の状態、すなわち初期位置では他移動体５０５が自移動体５０４に検知されるように設定される。
【０１１７】
円５０８の外側は自移動体５０４が他移動体５０５を検知できないので見失う範囲５０９となる。自移動体５０４が他移動体５０５を見失う範囲５０９に入った場合、自移動体５０４は見失いというシミュレーション結果となり、他移動体５０５の追跡という目的は達成されず、エピソードの種別は失敗となる。
【０１１８】
また自移動体５０４は、他移動体５０５に接近しすぎた場合には、他移動体５０５によって検知され、捕まるとする。この状態を拿捕として、追跡の継続という目的は達成されず、エピソードの種別は失敗となる。このために、自移動体５０４には拿捕範囲５１０が

設定され、円形の拿捕範囲５１０の内側に他移動体５０５が入ることで、拿捕を判定する。
【０１１９】
シミュレーションの開始から所定の、もしくは設定されたステップ数を経過するまで、見失い、及び拿捕の状態になることがなかった場合、自移動体５０４の追跡の目的は達成されたと判定され、エピソードの種別は成功となる。
【０１２０】
以上が図５に示した、追跡における行動環境の例の説明である。
【０１２１】
図６に、哨戒における自移動体、他移動体の行動環境の例を示す。
【０１２２】
哨戒領域６０１は予め設定される。自移動体６０２は哨戒領域６０１内を初期状態（位置、速度）から自由な方向６０５に移動できる。自移動体６０２の移動方向、速度はステップ毎の行動選択により変化する。
【０１２３】
自移動体６０２は他移動体を検知する能力を有し、検知距離６０３を半径とする円６０４の範囲に他移動体があれば、他移動体を検知したと判定する。
【０１２４】
他移動体は、図６の例では、垂直方向で哨戒領域６０１に侵入する他移動体６０６と、水平方向で侵入する他移動体６０７といったように行動にバリエーションがある。侵入位置は様々、もしくはランダムとする。侵入してきた他移動体６０６、６０７に対して、自移動体６０２が検知距離６０３に入れば検知に成功したと判定する。
【０１２５】
他移動体が検知されることなく哨戒領域６０１を通過したら検知に失敗したとする。自移動体６０２が他移動体を検知することが目的なので、検知成功ならプラスの報酬が付与され、検知失敗ならマイナスの報酬が与えられる。
【０１２６】
エピソードとしては、一つの他移動体の行動に対して検知の成功、もしくは失敗をもって、エピソードの種別を成功又は失敗に分類することができる。またエピソードのステップ数を予め定めておいて、また他移動体の数は多数として、強化学習してもよい。この場合、多くの報酬が得られるような自移動体６０２の行動が得られることとなる。
【０１２７】
他移動体の行動は、発生位置をランダムとすれば自移動体６０２の普遍的な哨戒行動が得られることとなるが、予め発生位置が指定されれば他移動体の特定の侵入の傾向に対する自移動体６０２の哨戒行動が得られることとなる。
【０１２８】
哨戒行動は、哨戒領域６０１の大きさに依存し、また他移動体の哨戒領域６０１への侵入行動パターンにも依存し、環境についての条件となる。
【０１２９】
図７Ａに哨戒領域と座標系の例を示す。図７Ｂに哨戒領域の構成と他移動体の行動パターンの例を示す。図７Ａでは、哨戒領域７０１を四角領域とし、中心（図心、重心）を原点として、Ｘ軸７０２、Ｙ軸７０３をとる。Ｘ方向の範囲Ｌｘ７０４、Ｙ方向の範囲Ｌｙ７０５により領域の大きさを設定する。範囲Ｌｘ７０４、Ｌｙ７０５の半分をｈＬｘ７０６、ｈＬｙ７０７とすれば、座標値としては±ｈＬｘ、±ｈＬｙとなる。
【０１３０】
図７Ｂに他移動体の行動パターンの例を示す。哨戒領域７０１に対して外側にオフセットを付加して広げた領域を他移動体発生領域７１０とする。他移動体は哨戒領域７０１の外部、かつ他移動体発生領域７１０の内部で発生する。
【０１３１】
哨戒領域７０１は四角形であり、他移動体は領域の４辺の上７１１、下７１２、右７１３、左７１４のそれぞれから移動方向７１５，７１６，７１７，７１８で侵入する。

【０１３２】
他移動体の哨戒領域７０１への侵入件数としては、予め設定されたステップ毎に１つずつ侵入させる、哨戒領域７０１内の他移動体の数を一定とする、などと決めることができる。
【０１３３】
以上が、図６、図７Ａ、図７Ｂに示した、哨戒における自移動体、他移動体の行動環境の説明である。
【０１３４】
以上が、本発明における移動体の行動の目的と環境についての説明である。
【０１３５】
＜強化学習＞
本発明における強化学習の内容について示す。強化学習は、図１、図３における強化学習部１０１及び図４の強化学習／エピソードデータ学習部４１０の処理に相当する。
【０１３６】
強化学習は、一般的にはマルコフ決定過程ＭＤＰ（Ｍａｒｋｏｖ Ｄｅｃｉｓｉｏｎ Ｐｒｏｃｅｓｓ）のモデルに基づいている。マルコフ決定過程は、ある状態において行動を選択して状態を更新し、この際の行動もしくは状態には報酬として表現される価値（又は状態価値）が伴う、というモデルである。
【０１３７】
強化学習の方式には、行動選択を行動価値のテーブル値により行うＱ学習、また行動価値関数を深層ニューラルネットワークとして行動価値から行動を選択するＤＱＮ（Ｄｅｅｐ Ｑ－ｌｅａｒｎｉｎｇ Ｎｅｔｗｏｒｋ）、方策（行動選択）関数と状態価値関数を深層ニューラルネットワークとしたＡ３Ｃ（Ａｓｙｎｃｈｒｏｎｏｕｓ Ａｄｖａｎｔａｇｅ Ａｃｔｏｒ－Ｃｒｉｔｉｃ）といった方法がある。
【０１３８】
深層ニューラルネットワークを利用したアルゴリズムは他にも存在するが、いずれも強化学習として、状態、行動、報酬によってモデルが訓練される。本発明ではＡ３Ｃを対象に説明する。
【０１３９】
強化学習は、エピソードの繰り返しにおいて、エージェントが状態に対して最適な行動を選択するように方策関数、状態価値関数を最適化する方法である。エージェントは学習の主体であって、例えば、本発明では自移動体のことである。エージェントは一つのエピソードにおいてステップ毎に、状態に対して行動を選択して、次の状態を得る。
【０１４０】
強化学習では、行動選択の結果や状態に対して報酬を与える。エピソードの終了条件を満たすならエピソードを終了する。エピソードを繰り返して、状態に対して良好な状態が得られるように正しく行動選択できるようになれば強化学習は終了となる。
【０１４１】
Ａ３ＣはＡｓｙｎｃｈｒｏｎｏｕｓ Ａｄｖａｎｔａｇｅ Ａｃｔｏｒ－Ｃｒｉｔｉｃの略であり、非同期処理、アドバンテージによる評価、アクター－クリティックとしての方策、状態価値関数モデルが特徴である。非同期処理は個別に用意された環境のそれぞれでエージェントのエピソード処理を行う。
【０１４２】
この場合、それぞれのエージェントが単一の方策関数と、状態価値関数を訓練することで、学習を効率化する工夫である。アドバンテージとは行動価値と状態価値の差であって、行動選択の良さを表す量である。アクター－クリティックとは、行動選択はアクターとして、状態価値はクリティックとして、別々のモデルで計算する方式であることを意味する。
【０１４３】
本発明での方策関数、状態価値関数の深層ニューラルネットワークによる構成の例を図

８に示す。深層ニューラルネットワークは状態量８０１を入力とし、行動８０２と、状態価値８０３を出力とする。

【０１４４】

状態量８０１から行動８０２までのネットワークが方策関数、状態量８０１から状態価値８４１までのネットワークが状態価値関数である。状態量８０１は、自移動体から他移動体までの距離８２１と、自移動体から見た他移動体の方位８２２と、自移動体の速度８２３である。

【０１４５】

行動８０２は自移動体の旋回行動の右８３１、左８３２、前８３３、変速行動の減速８３４、加速８３５である。状態価値８０３はそれ自体が一つの数値として状態価値８４１のみが項目となる。

【０１４６】

深層ニューラルネットワークは入力層８１１と、中間層８１２と、出力層８１３の多層構造として構成され、特に中間層８１２は複数の層となっている。なお図８ではノード（節点）を白丸で描いているが、全てのノードと層は描かずに点線で結合関係を表し、点線丸囲いでノードの集まりを表すこととして描いた。

【０１４７】

なお各ノードは関数であって、ノード同士を結ぶ線が入出力の関係を表す。状態量８０１は直接に入力層８１１に対応する。出力層８１３の各ノードは行動８０２、状態価値８０３の各項目に対応する。

【０１４８】

行動８０２の各項目に対応する出力層８１３のノードは総和が１となるように、確率の意味で事象分けされるようにＳｏｆｔｍａｘ関数が活性化関数として定義される。状態価値８４１に対応する出力関数にはＬｉｎｅａｒ関数で定義する。なお中間層８１２の各ノードにはＬｅｒｕ関数、Ｓｉｇｍｏｉｄ関数などが定義される。

【０１４９】

エピソードにおいてステップ毎のデータはメモリ１２に格納され、例えば１０ステップに１回といった、ステップ毎の所定のタイミングや、エピソード終了時に深層ニューラルネットワークが訓練される。各ステップのデータは状態量、行動、報酬である。

【０１５０】

図９Ａ、図９Ｂにエピソードにおける状態量、行動、報酬の一連のステップデータ（エピソードデータ）の構成を示す。図９Ａは失敗した場合のステップデータ９００Ａの例を示し、図９Ｂは成功した場合のステップデータ９００Ｂの例である。

【０１５１】

図９Ａ、図９Ｂの各図はそれぞれが１つのエピソードを構成するエピソードデータである。各図の項目は大きく分けるとステップ９０１と、状態量９０２と、行動９０３、報酬９０４、状態９０５である。状態９０５は深層ニューラルネットワークの訓練に利用しないが、参照、またエピソード分析で必要となる。

【０１５２】

状態量９０２は距離９１１、方位９１２、速度９１３であり、その数値は０以上１以下の範囲に正規化される。行動９０３は前９１４、右９１５、左９１６、減速９１７、加速９１８であり、一つのステップで選択される行動は一つであるため、一つの項目を１、他を０とする。なお、距離９１１は、自移動体から他移動体までの距離を示し、方位９１２は、自移動体から見た他移動体の方位を示し、速度９１３は、自移動体の速度を示す。

【０１５３】

報酬９０４はエピソードが失敗したら－１、成功したら１、それ以外のステップでは０とする。状態９０５は好位置、見失い、拿捕の３種類である。

【０１５４】

方策関数と状態価値関数の訓練の処理内容について説明する。深層ニューラルネットワークは状態量ｓを入力として、行動の出力である方策をπ（ｓ）、状態価値の出力をＶ（

ｓ）とするネットワークである。

【０１５５】

ネットワーク中のノードに対して定義される活性化関数のパラメータをθとする。方策は確率方策とも呼ばれ、状態ｓにおいて行動ａをとる確率π（ａ｜ｓ）とも記載する。方策関数、状態価値関数の訓練とは状態ｓ、行動ａ、次状態ｓ’、報酬ｒのデータから、それらの関係を予測する関係が得られるように、深層ニューラルネットワークのパラメータθを最適化することである。

【０１５６】

状態価値関数Ｖ（ｓ）は、方策π（ｓ）の下で以下の式（１）で表現される。

【０１５７】

【数１】

$$V(s) = E_{\pi(s)}[r + \gamma V(s')] \qquad (1)$$

【０１５８】

ここでＥは、添え字である方策π（ｓ）における期待値を意味する。割引率γは、次の状態価値（将来の価値）を現在の値に補正するための係数である。割引率γの値は、一例としては０．９９であるが、強化学習のための設定として調整できる。方策の価値は状態ｓの分布ρにおける期待値として以下の式（２）で表される。

【０１５９】

【数２】

$$J(\pi) = E_{\rho^s}[V(s)] \qquad (2)$$

【０１６０】

式（２）を割引済み報酬関数と呼ぶ。割引済み報酬関数の深層ニューラルネットワークのパラメータθに関する変化には、式（３）で表現される方策勾配定理がある。

【０１６１】

【数３】

$$\nabla_\theta J(\pi) = E_{s\sim\rho^\pi, a\sim\pi(s)}\left[A(s,a)\nabla_\theta \log\pi(s,a)\right] \qquad (3)$$

【０１６２】

ここで、∇は勾配（基底についての１階の偏微分）を意味する。期待値は状態ｓが方策πの分布ρの亘る範囲、行動ａは状態ｓに対する方策πに亘る範囲による。行動価値関数Ｑ（ｓ，ａ）、アドバンテージ関数Ａ（ｓ，ａ）は以下であり、データで計算可能である。

【０１６３】

【数４】

$$Q(s,a) = r + \gamma V(s') \qquad (4)$$

$$A(s,a) = Q(s,a) - V(s) \qquad (5)$$

【０１６４】

報酬関数を最大とするようにネットワークパラメータθを最適化すれば、状態に対して良い報酬が得られるような方策と状態価値の関係が得られる。報酬関数の負（マイナス）を方策損失とすれば、方策損失を最小化する。

【０１６５】
方策損失の他にも、行動価値と状態価値は一致していることが望ましいのでアドバンテージの絶対値の大きさを意味する価値損失も存在する。また状態に対して方策は一つに決まる方が望ましく、方策は確率的に決まるのでエントロピーでモデル化した正則化項も最適化に利用する。そこで損失関数を方策損失Ｌπ、価値損失Ｌｖ、正則化項Ｌｒｅｇにより定義し、損失関数を最小化する。
【０１６６】
【数５】

$$L = L_\pi + c_v L_v + c_{reg} L_{reg} \qquad (6)$$

【０１６７】
ここでｃｖ、ｃｒｅｇは係数である。
【０１６８】
方策損失は、割引済み報酬関数の定義から以下となる。ここで、訓練に使うｎはステップのデータ数である。
【０１６９】
【数６】

$$L_\pi = -J(\pi) = -\frac{1}{n}\sum_{i=1}^{n} A(s_i, a_i)\log\pi(a_i|s_i) \qquad (7)$$

【０１７０】
価値損失は、アドバンテージ関数の二乗とする。
【０１７１】
【数７】

$$L_v = \frac{1}{n}\sum_{i=1}^{n} A(s_0, a_0)_i^{\,2} \qquad (8)$$

【０１７２】
正則化項はエントロピーＨ（π（ｓ））を計算して得る。
【０１７３】
【数８】

$$H(\pi(s)) = -\sum_{k=1}^{n_{action}} \pi(s)_k \log\pi(s)_k \qquad (9)$$

$$L_{reg} = \sum_{i=1}^{n} H(\pi(s_i)) \qquad (10)$$

【０１７４】
ここでｎａｃｔｉｏｎは行動の数である。
【０１７５】
最適化計算によるニューラルネットワークの訓練には勾配報（ｇｒａｄｉｅｎｔ）を利用する。この訓練自体は深層学習における深層ニューラルネットワークの訓練と同じであ

る。
【０１７６】
以上が方策関数と状態価値関数の訓練の処理内容についての説明である。
【０１７７】
深層ニューラルネットワークの訓練を効率化するために、ステップ毎のデータのサンプリング方法について説明する。
【０１７８】
一般的にデータを用いた最適化計算においては、サンプルによる目的関数値の値が、最適化の目的から外れた値である方が、モデルのパラメータの調整幅が大きくなり、最適化の効率がよい。
【０１７９】
例えば、目的関数を最小化する場合、目的関数値が大きくなるデータを用いて最適化計算をすることで、目的関数の値を大幅に小さくするようにパラメータの値を調整できる。目的関数値の変化を大きくできることで、パラメータの調整幅を大きくできるのである。
【０１８０】
目的関数は式（６）で表現される損失関数であり、この最小化が最適化の目的である。その項となっている方策関数、価値関数、正則化項は式（７）、（８）、（９）に示されるように、状態量ｓと行動ａのデータを利用して計算されることを示す。これらの計算には方策関数π（ｓ）もしくはπ（ａ｜ｓ）と、アドバンテージ関数Ａ（ｓ，ａ）が利用される。
【０１８１】
方策関数は深層ニューラルネットワークで状態量から行動を選択する関数であり、状態量ｓから行動別の値を取得できる。アドバンテージ関数は式（５）で定義され、また式（４）の行動価値関数Ｑ（ｓ，ａ）を利用する。
【０１８２】
行動価値関数Ｑ（ｓ，ａ）は式（４）から状態価値関数Ｖ（ｓ’）と報酬ｒで計算される。状態価値関数は深層ニューラルネットワークで状態量から状態価値を算出する関数であり、状態量ｓで値を取得できる。
【０１８３】
行動価値関数においては、次状態ｓ’はエピソードデータでのステップで、次の行のステップの状態量ｓを用いて計算する。最終行のステップについては状態価値関数Ｖ（ｓ’）をゼロとすればよい。以上から損失関数値は、ステップ別に状態量ｓ、行動ａ、報酬ｒから計算できる。
【０１８４】
個々のステップに対して損失関数値が得られるので、損失関数値が大きくなる順にステップデータを並べて、大きな値となったステップの状態量、行動、報酬データをサンプルすることで、訓練を効率化できる訓練用のデータが得られる。
【０１８５】
多くのステップ毎のデータから訓練に必要な数のデータを損失関数値の大きさの順にサンプリングしても、また限られたステップデータから重複も許して確率的にサンプリングしてもよい。また損失関数値が小さいことも、それまでに訓練されてきた深層ニューラルネットワークのパラメータの特徴を反映しているので損失関数値が小さくなるステップデータも混ぜてサンプリングしてもよい。
【０１８６】
以上が深層ニューラルネットワークの訓練を効率化するための、ステップデータのサンプリング方法についての説明である。処理の主体は行動制御計画装置１０５である。
【０１８７】
一連の強化学習の処理手順の一例を、図１０のフローチャートにより説明する。
【０１８８】
ステップ１００１からステップ１０１８はエピソードの繰り返しである。

【０１８９】
エピソードの処理においては、まず環境、自移動体、他移動体の状態をリセットする１００２。
【０１９０】
ステップ１００３からステップ１０１４は、１つのエピソードにおけるステップの繰り返しである。
【０１９１】
強化学習／エピソードデータ学習部４１０は、自移動体の行動を選択する（１００４）。この処理は、方策関数に現在の状態量を入力することで行われる。
【０１９２】
そして強化学習／エピソードデータ学習部４１０は選択された行動を実行し、状態量を更新する（１００５）。強化学習／エピソードデータ学習部４１０は、シミュレータ４００の機能を利用して状態量を算出する。
【０１９３】
強化学習／エピソードデータ学習部４１０は、自移動体の状態量に基づき、エピソードの失敗に関するＮＧ判定を行う（１００６）。追跡の例ならば、見失い、拿捕の判定が失敗に相当する。
【０１９４】
強化学習／エピソードデータ学習部４１０は、ＮＧ判定１００６がｙｅｓ（ＮＧである）ならば、エピソード失敗のフラグを設定する（１００７）。
【０１９５】
次に、強化学習／エピソードデータ学習部４１０は、エピソードで実施したステップ数の判定を行う（１００８）。この判定は、実施したステップ数が所定数を超えていれば、エピソードの終了と判定してステップ１００９へ進み、そうでない場合にはステップ１０１１に進む。
なお、上記ステップ１００７で、既に失敗フラグが設定されている場合は、ステップ１００８の判定自体が不要であり、判定結果としてはｎｏ（終了ステップ数ではない）となる。またその場合にはエピソード終了フラグも設定する。
【０１９６】
ステップ数の判定（１００８）でｙｅｓ（処理したステップ数が所定の終了ステップ数である）の場合、強化学習／エピソードデータ学習部４１０は、成功フラグを設定する（１００９）。これは、失敗ではなく、指定されたステップ数の間、処理を継続できたのでエピソードの種別が成功ということを意味する。処理が追跡の例ならば、好位置の状態を継続してきたことが成功に相当する。
【０１９７】
なお、哨戒の例については、ＮＧ判定（１００６）でＯＫ（ｎｏ）の場合には、他移動体の検知として、検知数をカウントする。ＮＧ（ｙｅｓ）の場合には検知失敗の見失いと判定して見失い数をカウントする。いずれの場合も他移動体を環境から除外する処理を行う。ステップ数判定（１００８）は、終了ステップ数だけの判定になり、特に成功フラグ設定（１００９）は不要である。
【０１９８】
ステップ数判定（１００８）でｙｅｓの場合には、強化学習／エピソードデータ学習部４１０がエピソード終了フラグを設定する（１０１０）。
【０１９９】
強化学習／エピソードデータ学習部４１０は、深層ニューラルネットワークの訓練のため、訓練データを設定する（１０１１）。このステップ１０１１は、訓練のためにメモリ１２から一纏まりのデータを準備することであって、訓練を効率化するためのサンプリングといった処理も含まれる。
【０２００】
そして、強化学習／エピソードデータ学習部４１０は、上記ステップ１０１１で設定し

た訓練データを深層ニューラルネットワークへ入力して訓練する（１０１２）。
【０２０１】
ステップ１０１１、１０１２の処理は、必ず１ステップ毎に行う必要はなく、数ステップ毎、またエピソード終了時といったタイミングで実施するように、条件分岐してもよい。
【０２０２】
次に、強化学習／エピソードデータ学習部４１０がエピソードの終了を判定する（１０１３）。強化学習／エピソードデータ学習部４１０は、エピソード終了フラグが設定されていればエピソードの終了値を判定する。判定がｎｏ（エピソード終了フラグの設定無し）の場合、ステップ１００３に戻って上記処理を繰り返す。
【０２０３】
エピソード終了判定１０１３でｙｅｓ（終了フラグの設定あり）の場合、ステップ１００３からステップ１０１４のループを抜けて、ステップ１０１５へ進む。
【０２０４】
強化学習／エピソードデータ学習部４１０は、エピソード内のステップの繰り返しを終了したら、エピソードデータを出力してエピソードデータベース１２２に蓄積する（１０１５）。
【０２０５】
次に強化学習／エピソードデータ学習部４１０は、エピソードの連続成功回数を判定する（１０１６）。強化学習／エピソードデータ学習部４１０は、エピソードが１回成功して強化学習を終了するなら、エピソード成功判定はｙｅｓでよい。例えば１０回連続成功を終了条件とした場合、１０回に満たないなら判定結果はｎｏであるが、連続成功回数に１を加算してステップ１０１７に進む。エピソードが失敗ならば連続成功回数は０として、判定結果をｎｏとしてステップ１０１７に進む。
【０２０６】
次に強化学習／エピソードデータ学習部４１０は、実施したエピソードの回数を判定する（１０１７）。強化学習はエピソード毎の繰り返しであり、エピソードが成功回数判定（１０１６）でｙｅｓにならない限り、無限にステップ１００１からステップ１０１８を繰り返すこととなる。そこで、強化学習／エピソードデータ学習部４１０は、エピソード回数の上限値を設け、エピソードの実施回数が上限値を超える場合にｙｅｓと判定してエピソード繰り返しのループを抜ける。この場合、強化学習は失敗となる。
【０２０７】
最後に強化学習の終了として、学習終了の処理を行う（１０１９）。強化学習／エピソードデータ学習部４１０は、強化学習過程のエピソードの情報などのログを出力する。
【０２０８】
以上が、図１０のフローチャートによる、強化学習の一連の処理手順の一例の説明である。以上が、本発明における強化学習の内容についての説明である。
【０２０９】
＜エピソード分析＞
本発明における、エピソード分析の内容についての例を示す。
【０２１０】
エピソード分析は、図１、図３におけるエピソード分析部１０２及び図４のエピソード分析部４３０の内容に相当する。
【０２１１】
エピソード分析とは、エピソードデータベース４１９に蓄積されたエピソードデータを、強化学習過程におけるエピソードの成功、失敗の順を、他移動体の行動パターンについて個別に分析することで、学習パターンを構成する、という一連の処理である。
【０２１２】
個別のエピソードデータは図９Ａ、図９Ｂに示した、１つのエピソードを構成するステップ９０１毎の状態量９０２と、行動９０３と、報酬９０４及び、状態９０５のデータである。１回の強化学習の結果として、強化学習過程について、１つ以上の、実際には多数

のエピソードデータがエピソードデータベース４１９に格納される。
【０２１３】
以降の説明におけるエピソードデータについては、艦船による追跡の強化学習についてのデータの例である。
【０２１４】
１つの強化学習過程についてのエピソードデータを図１１Ａ、図１１Ｂに示す。図１１Ａは強化学習によってエピソードが成功した例を示し、図１１Ｂは強化学習によってエピソードが失敗した例である。図示のグラフは横軸がエピソード（数）１１０１、縦軸がエピソード終了時のステップ数１１０２である。プロットの形は、凡例１１０３に示すように、見失いが黒四角（塗りつぶし四角）、拿捕が白三角（白抜き三角）、好位置が白丸（白抜き丸）である。また、シミュレーション結果が見失いと、拿捕の場合はエピソードの種別が失敗となり、シミュレーション結果が好位置の場合はエピソードの種別が失敗となる。
【０２１５】
図１１Ａのグラフでは、強化学習過程において１５０エピソード程度までは、ほぼ１００ステップを超えずに見失いで、エピソードが失敗している。また、多少の拿捕のケースも見られる。１５０エピソード以降、エピソードは失敗であるが、ステップ数は増え始める。
【０２１６】
２００エピソードを超えると、ステップ数が４００ステップ達成の好位置でのエピソードの成功が見られるようになる。２５０エピソード以降はほぼエピソードの種別は成功であり、エピソードの成功が繰り返されて２７０エピソードで強化学習は成功として学習は完了した。
【０２１７】
失敗の例として図１１Ｂのグラフでは、１００エピソードまではほぼ１００ステップを超えていないが、１００エピソード以降、２００エピソードまでで成功のエピソードが出現した。２００エピソードから３００エピソードの間は成功エピソードは多いが、シミュレーション結果が拿捕となった失敗エピソードも多い。
【０２１８】
３００エピソード付近からは拿捕ばかりとなり、３５０エピソード以降、拿捕と見失いによる失敗エピソードばかりとなった。成功エピソードが連続することはなく、５５０エピソードで強化学習は失敗となった。
【０２１９】
他移動体の同一のシミュレーション結果で、エピソードが成功した強化学習の強化学習過程についての、エピソード失敗、成功別のステップ数のプロットを図１２に示す。なお、以下の図では、上記図１１Ａと同様に、シミュレーション結果が見失いと、拿捕の場合はエピソードの種別が失敗となり、シミュレーション結果が好位置の場合はエピソードの種別が失敗となる。
【０２２０】
図１２では、４００ステップ達成のエピソード成功のプロットはエピソードの数が１５０から３５０、平均的には２５０エピソードで学習が完了している。
【０２２１】
見失い、又は拿捕で失敗したエピソードのステップ数も、強化学習過程のエピソード数に沿って徐々に多くなっている。但し１００エピソードまでは１５０ステップ、平均的に見ればおよそ５０ステップ程度までの増え方である。
【０２２２】
なお、プロットの状況を見ると強化学習の成功までの過程において、エピソードの種別のばらつきが大きい。強化学習では行動選択は乱択であり、深層ニューラルネットワーク１０の訓練の結果にばらつきが大きく入るためと考えられる。理想的なエピソードを経た訓練ならば、エピソード数は少なくできると考えられる。

【０２２３】
学習パターンを検討するにあたり、見失いと、拿捕の失敗エピソードと４００ステップ連続好位置の成功エピソードを、エピソード数とステップ数による単純な線図として、図１３に示すように整理する。図中線１３０１は見失い、破線１３０２は拿捕を表す。これら２つの線は同じ傾向であるとした。網掛け線１３０３は成功エピソードである。図１２における成功、失敗のプロットの平均的な値で描いた。
【０２２４】
強化学習を行った深層ニューラルネットワーク１０がシミュレーションに成功する場合の、エピソード数の観点で理想的な、強化学習過程における失敗、成功エピソードの出現の範囲を図１４に示す。図中、黒塗りつぶし１４０１は見失いを示し、斜め網掛け１４０２は拿捕を示し、点網掛け１４０３は連続４００ステップで好位置を示す。
【０２２５】
エピソードデータ学習部４２０（１２２）は、１５０エピソードで強化学習が完了すると想定して、１５０エピソードまでの４０エピソード分を成功のエピソードに割り当てる。強化学習過程の初期においては少ないステップ数での見失い、拿捕の失敗が多いので、ステップ数０から１００の範囲で、２０エピソードずつ見失い、拿捕の失敗のエピソードに割り当てる。継いでステップ数１００から２００の範囲で、２０エピソードずつ見失い、拿捕の失敗を割り当てる。８０エピソード目から成功のエピソードまでのエピソード数１１０の範囲を、ステップ数２００から３００、ステップ数３００から４００の見失い、拿捕のエピソードに割り当てる。
【０２２６】
なお、エピソードデータ学習部４２０（１２２）が、上記ステップ数の範囲とエピソード数の範囲に割り当てるシミュレーション結果の種別（見失い、拿捕、好位置）は、予め設定した比率（又は数）で割り当てることができる。
【０２２７】
図１４に示した強化学習過程における成功、失敗のエピソードと、それらのステップ数の範囲より、図１５に示す学習パターンが得られる。学習パターン１５００は、エピソード結果１５０１と、終了ステップ数範囲の下限１５０２及び上限１５０３と、エピソード数１５０４で定義されるデータ構造である。エピソード結果１５０１はエピソード終了時の状態である。終了ステップ数範囲は下限１５０２と、上限１５０３で指定される。
【０２２８】
図１５に示した学習パターン１５００は、図１６のグラフで示すように横軸にエピソード数、縦軸にステップ数としてプロットされる。図１６では、ステップ数範囲は範囲の中心の値としている。
【０２２９】
以上により、他移動体の行動バターン（拿捕、見失い、好位置）に対して、エピソードデータベース４１９から収集したエピソードデータを用いて学習パターンを構成する例を示した。図４の行動制御計画装置１０５におけるエピソード分析部４３０は、エピソードデータ収集機能４３１、ステップ毎エピソード集計機能４３３、学習パターン構成機能４３４により処理される。
【０２３０】
エピソード分析は、図１２の全てのエピソードデータの平均やばらつきなどを評価することにより、学習パターンの構成要素である、強化学習開始からのエピソード数に対するステップ数の範囲を決める処理であって、計算処理によって自動化できる。
【０２３１】
学習パターンを用いるエピソードデータ学習において、行動制御計画装置１０５はエピソードデータベース４１９からエピソードデータを収集する。以降でエピソードデータの収集方法について説明する。なお、エピソードデータの収集は他移動体のシミュレーション結果別に行う。
【０２３２】

この処理は、図３のフローチャートにおけるエピソード収集（学習用）３１１の内容であり、また図４の行動制御計画装置１０５の構成のエピソードデータ収集・一覧化機能４２２による処理である。

【０２３３】

図１５に示した学習パターンに対応するエピソードデータのサンプリング結果を図１７に示す。図１７は、エピソードの結果順に、エピソード数の範囲毎にステップ数の範囲内のエピソードデータのサンプルが入っていることを示す。

【０２３４】

例えば、１エピソードから２０エピソードまではステップ数１００以下の見失い（図中四角）のサンプルを示し、２１エピソードから４０エピソードまではステップ数１００以下の拿捕（図中三角）のサンプルを示し、また、１１１エピソードから１４０エピソードまではステップ数４００以下の成功のエピソードのサンプル、となっている。

【０２３５】

図１７の強化学習過程のエピソードの見失い、拿捕、好位置の分布は、図１１Ａに示した強化学習の成功例の分布とかなり異なっている。図１７は見失い、拿捕、好位置のシミュレーション結果について、横軸（エピソード数）方向のエピソード範囲でのばらつきがないためである。

【０２３６】

そこで、図１７のような学習パターンによるサンプリングにエピソード範囲でのばらつきを与えるために、強化学習過程におけるエピソード（の位置の）間の置換をする方法がある。エピソード間の置換とは、例えば、強化学習開始からの１０エピソード目と２０エピソード目のエピソードを交換する操作である。置換対象のエピソードの選択をランダムで決める置換をランダム置換と呼ぶ。

【０２３７】

図１８Ａに置換数５０のランダム置換の例を示し、図１８Ｂに置換数１００のランダム置換の例を示し、強化学習過程についてのエピソード数とステップ数の分布を示す。図１８Ａ、図１８Ｂでは、１３０エピソード以降のエピソードは置換の対象とはしていない。このように、学習パターンによるサンプリングにエピソード範囲でばらつきを与えることができる。なお、置換という言葉で説明したが、シャッフルと言ってもよい。

【０２３８】

以上が学習パターンによるエピソードデータ収集方法についての説明である。

【０２３９】

以上が、本発明における、エピソード分析の内容についての例の説明である。

【０２４０】

＜エピソードデータ学習＞

本発明における、エピソードデータ学習の内容についての例を示す。

【０２４１】

エピソードデータ学習は、図１、図３におけるエピソードデータ学習部１０３及び図４のエピソードデータ学習部４１０の内容に相当する。

【０２４２】

エピソードデータ学習とは、他移動体のシミュレーション結果別の学習パターンに従ってエピソードデータを収集し、エピソード毎にエピソードのステップデータ（エピソードデータ）を直接に用いて深層ニューラルネットワークを訓練することで、状況に応じた行動の選択を学習する方法である。

【０２４３】

エピソードデータ学習にあたっては、まず条件を取得することは強化学習と同様である。但し、エピソードデータ学習のためには、他移動体のシミュレーション結果は複数であって、それらに対応する他移動体の条件、また学習パターンを取得することになる。

【０２４４】

エピソードデータ収集については、本実施例の学習パターンに対するエピソードデータ

の収集について説明した通りである。
【０２４５】
エピソードデータ学習では、エピソードデータを直接に利用して、エピソードにおけるステップのデータを用いて深層ニューラルネットワークを訓練する。エピソードデータ学習の目的は、状態量に応じて行動を選択することであり、学習のモデルは、強化学習と同様にマルコフ決定過程ＭＤＰのモデルに基づいている。
【０２４６】
つまり上記強化学習の内容で示したように、方策関数と状態価値関数を深層ニューラルネットワークで定義し、エピソードの繰り返しで、各エピソードではステップ毎のデータを用いて深層ニューラルネットワークを訓練するものである。
【０２４７】
深層ニューラルネットワークの構成、また訓練のための、例えば、上記（１）式から（１０）式で表現される計算手法を用いる。また、ステップデータのサンプリングについても強化学習と同じである。
【０２４８】
エピソードデータ学習と強化学習の違いは、強化学習はステップ毎にエージェントの行動選択に乱択と方策関数を利用した選択を混在させ、行動により状態を更新することにより状態量、行動、報酬のデータを得る。
【０２４９】
これに対して、エピソードデータ学習は状態量、行動、報酬のデータを、例えばファイルから読み込むなど、直接に取得することである。エピソードデータ学習には、強化学習における探索と活用（ｅｘｐｌｏｒａｔｉｏｎ ａｎｄ ｅｘｐｌｏｉｔａｔｉｏｎ）といった特長はない。
【０２５０】
この点でエピソードデータ学習そのものには、シミュレータや、場合によっては現実の環境における設備や実験は不要である。但し、学習が成功しない、無意味なエピソードデータは利用せず、事前に強化学習した際に得たエピソードデータを利用する。
【０２５１】
ここで学習に必要となるエピソードデータを学習パターンとして選別することで、学習を効率化し、さらに学習が難しくなる異なる環境や、他移動体のシミュレーション結果に対する学習を実現する手法である。
【０２５２】
一連のエピソードデータ学習の処理手順の一例を、図１９のフローチャートにより説明する。図１９における処理の主体は行動制御計画装置１０５である。
【０２５３】
まず、エピソードデータ学習部４１０は、エピソードデータをエピソードデータベース４１９（１２２）から収集する（１９０１）。
【０２５４】
ステップ１９０２からステップ１９１８までの範囲はエピソード毎の繰り返しである。エピソードの繰り返しは学習パターンの指定に基づいて、エピソードデータベース１２２から収集されたエピソードデータを、強化学習過程における順序での繰り返しである。
【０２５５】
なお、このエピソードデータには、他移動体の複数のシミュレーション結果のエピソードデータが混在している。例えば、シミュレーション結果が２つで、それぞれエピソードの数が同じである場合には、シミュレーション結果毎に一つずつのエピソードを並べる。１つのシミュレーション結果のエピソードの数が、もう一つの倍である場合、倍の方のエピソードの２つに対し、もう一つのエピソードを１つ並べるといった形での混在である。
【０２５６】
エピソードの繰り返しにおいては、エピソードデータ学習部４１０は、まず環境、自移動体、他移動体の状態をリセットする（１９０３）。

【０２５７】
ステップ１９０４からステップ１９０９の範囲は１エピソードにおけるステップ毎の繰り返しである。
【０２５８】
エピソードデータ学習部４１０は、現在のステップの状態量、行動、報酬を取得する（１９０５）。強化学習では、図１０に示したように、ステップの繰り返しでは、行動選択、行動実行、状態に基づく失敗、成功の判定があったが、エピソードデータ学習では、ステップのデータを読み込むだけである。
【０２５９】
次にエピソードデータ学習部４１０は、ＤＮＮ訓練機能４１４を利用してステップのデータで深層ニューラルネットワークを訓練する（１９０６）。
【０２６０】
エピソードデータ学習部４１０は、エピソードの終了を判定する（１９０７）。エピソードデータでステップ数が決まっているので最後のステップであればエピソードの終了となる。一方、終了でなければステップを更新し（１９０８）、次のステップで上記処理を繰り返す。
【０２６１】
エピソードデータ学習部４１０は、エピソードの処理が終了する度に、学習の完了のテストをステップ１９１０～１９１５で行う。
【０２６２】
ステップ１９１０からステップ１９１５の範囲は、全てのシミュレーション結果についての繰り返しである。
【０２６３】
そして、ステップ１９１１からステップ１９１４の範囲は、指定された試行回数についての繰り返しである。この試行回数が、例えば５回である場合、ある一つのシミュレーション結果について５回の試行を行うということであり、連続して５回成功ならば、そのシミュレーション結果に対する学習は成功したということを意味する。
【０２６４】
まず、エピソードデータ学習部４１０は、訓練結果（学習結果）の深層ニューラルネットワークを用いて指定されたエピソードを実行する（１９１２）。これは図３におけるステップ３０５と同様の処理であって、初期状態から１ステップ毎に方策関数により行動を選択し、状態を更新することである。
【０２６５】
次に、エピソードデータ学習部４１０は、エピソードの失敗か成功かを判定する（１９１３）。失敗した場合がｙｅｓであって、直ちにステップ１９１０からステップ１９１５の繰り返しを抜ける。失敗しない場合（ｎｏ）には、繰り返しを継続する。全ての繰り返しで成功したなら、ステップ１９１３からステップ１９１０の繰り返しを抜けることはない。
【０２６６】
次に、エピソードデータ学習部４１０は、全てのシミュレーション結果で成功したかを判定する（１９１６）。成功した場合がｙｅｓであり、この場合には学習が成功でありエピソードの繰り返しから抜ける。
【０２６７】
次にエピソード回数を判定する（１９１７）。この処理は、所定のエピソード繰り返し回数の間にエピソードが成功しない場合の判定であり、判定がｙｅｓの場合、学習失敗としてエピソードの繰り返しから抜ける。全てのエピソードの繰り返しで学習を行う場合には、ステップ１９１７の判定は不要である。
【０２６８】
最後にエピソードデータ学習部４１０は、学習終了の処理を行う（１９１９）。この処理は、ログなどの出力処理である。

【０２６９】
以上が、図１９のフローチャートによる、一連のエピソードデータ学習の処理手順の例の説明である。上記処理によって、他移動体の複数のシミュレーション結果（例えば、拿捕、見失い、好位置など）について予め設定された学習パターンに一致するエピソードデータを抽出してエピソードデータ学習を行って、学習結果をエピソードデータベース１２２へ蓄積することができる。
【０２７０】
これにより、強化学習で算出された全てのエピソードデータベース１２２を用いるのではなく、予め設定された学習パターンで抽出したエピソードデータをエピソードデータベース１２２から抽出することで、複数のシミュレーション結果のエピソードデータで深層ニューラルネットワーク１０を直接訓練する。
【０２７１】
エピソードデータ学習では学習パターンの設定に応じて、効率的に深層ニューラルネットワークの訓練が可能となり、例えば、少ないエピソード数で訓練を完了させながらも、学習が成功しやすくなる。これにより、深層ニューラルネットワーク１０の行動制御の精度を向上させて、自律的に行動する移動体の運用を容易にすることが可能となる。
【０２７２】
以上が、本発明におけるエピソードデータ学習の内容についての例の説明である。
【０２７３】
本発明の移動体の行動制御計画装置は、計算処理を主な技術としており、ソフトウェアを主な実現手段とする。したがってパッケージソフト、クラウドサービスといった形態での実現が考えられる。学習した結果の方策関数、状態価値関数は制御ソフトウェアとして機械に組み込める。よって、実際に行動する艦艇、車両といった移動体に組み込まれたり、また移動体を遠隔的に制御する他の移動体や装置、施設での活用も考えられる。本発明はそのような活用の範囲を限定するものではない。
【０２７４】
以上が、他移動体の様々な行動に対する自移動体の行動制御のために、他移動体の行動に対して強化学習を行い、その結果の様々な行動のエピソードデータを用いたエピソードデータ学習によって、自移動体の行動制御を計画する技術の説明である。
【０２７５】
＜結び＞
以上説明したように、本発明の行動制御計画装置１０５は、自移動体に対して他移動体の行動が多様な関係となる場合でも、例えば追跡、検知といった目的を達成するための、自移動体の行動制御を深層ニューラルネットワークで学習し、学習結果の深層ニューラルネットワークを自律的に行動可能な移動体の制御に適用することが可能となる。
【０２７６】
また、エピソード中のステップのデータを、深層ニューラルネットワークの訓練におけるパラメータの最適化のため、目的関数である損失関数値に基づきエピソードデータに優先付けすることで、効率的に強化学習でき、強化学習を行った深層ニューラルネットワークではエピソードが成功する可能性を向上させることが可能となる。
【０２７７】
個別の環境、特に他移動体のシミュレーション結果に対しては強化学習を実施した深層ニューラルネットワーク１０がシミュレーションで成功する。成功した強化学習のエピソードをエピソードデータとして蓄積する。複数の他移動体のシミュレーション結果のエピソードデータを用いて、エピソードのステップデータで直接に深層ニューラルネットワークの訓練を行うエピソードデータ学習により、複数のシミュレーション結果に対応する行動制御を深層ニューラルネットワークに学習させることができる。
【０２７８】
エピソードデータ学習の成功のために、蓄積されたエピソードデータを収集して分析する。個別の環境、他移動体のシミュレーション結果の種類別に、成功した強化学習につい

て、エピソードの順序に関して成功と失敗のエピソードとそのステップ数を一覧化することで、エピソード順序に対して失敗、成功とステップ数の範囲を学習パターンとして定義する。

【０２７９】

上記定義されたエピソード順序に対するエピソードデータをエピソードデータ学習に利用することでエピソードデータ学習を実施することで、自移動体に対して他移動体の行動が多様な関係となる場合でもエピソードが成功する可能性を向上させることができる。

【０２８０】

強化学習、エピソードデータ学習により移動体の行動制御を計画すれば、マニュアルによる計画が困難な多様な環境、他移動体の行動パターンに対応することが可能となる。さらには、本実施例の行動制御計画装置は、人間が考え付くことが困難と考えられる行動制御内容が得られる可能性がある。

【０２８１】

なお、上記実施例において、行動制御計画装置１０５で学習した深層ニューラルネットワークを適用する移動体として哨戒又は追跡を行う艦船の行動制御に本発明を適用する例を示したが、これに限定されるものではなく、航空機や車両、潜水艦或いはドローン等で自律的に行動する移動体に適用することができる。

【０２８２】

また、上記実施例において、行動制御計画装置１０５は、シミュレータ４００（又は１１２）を含む例を示したが、これに限定されるものではなく、シミュレータ４００を外部の計算機で実施してもよい。この場合、行動制御計画装置１０５は、図１の条件設定１１１で受け付けたシミュレーション条件を、外部のシミュレータに送信してシミュレーションを実行させ、シミュレーションの結果を取得する。この例では、行動制御計画装置１０５と外部シミュレータで構成された行動制御計画システムとすることができる。そして、行動制御計画装置１０５は、外部からのシミュレーション結果に基づいて深層ニューラルネットワーク１０の強化学習を実施する。

【０２８３】

以上のように、上記実施例の行動制御計画装置１０５は、以下のような構成とすることができる。

【０２８４】

（１）プロセッサ（１１）とメモリ（１２）を有して、他移動体の行動に対して自移動体の行動を計画する行動制御計画装置（１０５）であって、前記自移動体と他移動体の行動をシミュレーション条件として設定するシミュレーション条件設定部（１１１）と、前記シミュレーション条件と予め設定された機械学習モデル（深層ニューラルネットワーク１０）に基づいて、前記自移動体（５０４）と他移動体（５０５）の位置関係を所定の時間間隔毎にステップデータ（９００Ａ）として出力するシミュレータ（１１２）と、前記ステップデータ（９００Ａ）を取得して、前記機械学習モデル（１０）に前記ステップデータ（９００Ａ）を与えて学習結果として前記自移動体と他移動体の前記位置関係を示す状態量（９０２）と報酬（９０４）と状態（９０５）を含むエピソードデータを出力させ、当該エピソードデータをエピソードデータ蓄積部（エピソードデータベース１２２）に蓄積する強化学習部（１０１）と、前記エピソードデータ蓄積部（１２２）に蓄積された前記エピソードデータを分析して学習パターンを生成するエピソード分析部（１０２）と、前記学習パターンに該当する前記エピソードデータを前記エピソードデータ蓄積部（１２２）から取得して学習用エピソードデータを生成（エピソードデータ収集・一覧化機能４２２）し、前記機械学習モデル（１０）に前記学習用エピソードデータを与えて学習させるエピソードデータ学習部（１０３）と、を有することを特徴とする行動制御計画装置。

【０２８５】

上記構成により、行動制御計画装置１０５は、様々な環境に対して、特に自移動体に対して他移動体の行動が多様な関係となる場合でも、例えば追跡、検知といった目的を達成するための、自移動体の行動を学習できる。

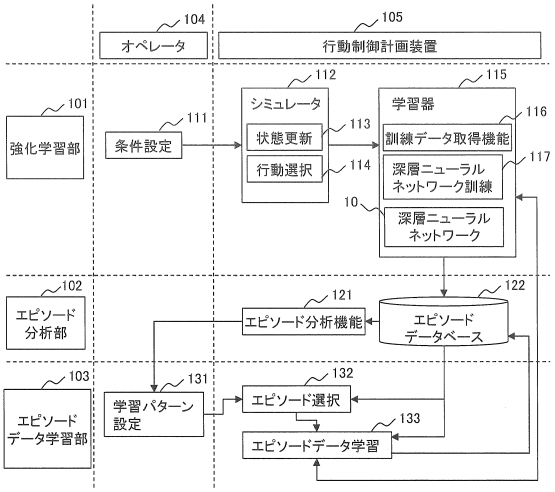
【０２８６】
（２）上記（１）に記載の行動制御計画装置（１０５）であって、前記強化学習部（１０１）は、前記ステップデータ（９００Ａ）の開始から所定の終了条件に達するまでのステップデータ（９００Ａ）を１つのエピソードとし、前記エピソードは前記ステップデータ（９００Ａ）に対応するステップを前記エピソードデータに設定し、前記エピソードデータの前記状態（９０５）には前記ステップデータ（９００Ａ）毎のシミュレーション結果を設定し、前記エピソードには前記シミュレーション結果に応じて所定の目的を達成したか否かを示すエピソード種別（１００７、１００９）を設定し、前記エピソード分析部（１０２）は、前記エピソード種別と前記シミュレーション結果と前記エピソードの数と前記エピソードデータのステップ数に基づいて前記学習パターンを生成することを特徴とする行動制御計画装置。
【０２８７】
上記構成により、エピソード分析部１０２は、全てのエピソードデータの平均やばらつきなどを評価することにより、学習パターンの構成要素である、強化学習開始からのエピソード数に対するステップ数の範囲を絞り込むことで、学習対象の数を抑制しながら効率よく機械学習モデルの学習（訓練）を実現することが可能となる。
【０２８８】
（３）上記（２）に記載の行動制御計画装置（１０５）であって、前記エピソード分析部（１０２）は、前記シミュレーション結果毎に前記エピソードの数の範囲（１５０４）と、前記エピソードデータのステップ数の範囲（１５０２、１５０３）を設定し、前記エピソードデータ学習部（１０３）は、前記エピソードデータ蓄積部（１２２）から、前記シミュレーション結果毎に前記エピソードの数の範囲（１５０４）と、前記エピソードデータのステップ数の範囲（１５０２、１５０３）に該当するエピソードを取得して前記機械学習モデル（１０）に学習させることを特徴とする行動制御計画装置。
【０２８９】
上記構成により、エピソードデータ学習部１０３は、学習パターンの設定に応じて、効率的に深層ニューラルネットワーク１０の訓練が可能となり、例えば、少ないエピソード数で訓練を完了させながらも、学習が成功しやすくなる。これにより、深層ニューラルネットワーク１０の行動制御の精度を向上させて、自律的に行動する移動体の運用を容易にすることが可能となる。
【０２９０】
（４）上記（３）に記載の行動制御計画装置（１０５）であって、前記エピソードデータ学習部（１０３）は、前記シミュレーション結果毎に、前記エピソードの数（１５０４）とステップ数の範囲（１５０２、１５０３）で抽出したエピソードを、前記エピソードの順序で所定の置換を行うことを特徴とする行動制御計画装置。
上記構成により、エピソード分析部１０２は、エピソードの順序で所定の置換を行うことで、深層ニューラルネットワーク１０の訓練結果のばらつきを抑制することができる。
【０２９１】
（５）上記（３）に記載の行動制御計画装置（１０５）であって、前記エピソードデータ学習部（１０３）は、前記取得したエピソードに対応する前記エピソードデータを前記機械学習モデル（１０）に直接入力して学習させる（１９０５）ことを特徴とする行動制御計画装置。
【０２９２】
上記構成により、エピソードデータ学習部１０３は、エピソードデータを直接用いて深層ニューラルネットワーク１０を訓練することで、状況に応じた行動の選択を学習することが可能となる。
【０２９３】
（６）上記（１）に記載の行動制御計画装置（１０５）であって、前記所定の目的は、前記自移動体（５０４）が前記他移動体（５０５）の追跡であることを特徴とする行動制御計画装置。

【０２９４】
上記構成により、自移動体５０４が他移動体５０５を見失うことなく、かつ、他移動体５０５に拿捕されることなく追跡を行う深層ニューラルネットワーク１０を生成することができる。
【０２９５】
（７）上記（２）に記載の行動制御計画装置（１０５）であって、前記エピソード分析部（１０２）は、前記エピソードの数の範囲（１５０４）と、前記エピソードデータのステップ数の範囲（１５０２、１５０３）において、前記シミュレーション結果の種別に応じた所定の比率で前記エピソードを割り当てることを特徴とする行動制御計画装置。
【０２９６】
エピソード分析部１０２は、エピソード数の範囲と、エピソードデータのステップ数の範囲内で、シミュレーション結果の種別毎に割り当てる比率を設定することで、エピソード数とステップ数に応じたシミュレーション結果の種別の傾向を適用することが可能となる。
【０２９７】
なお、本発明は上記した実施例に限定されるものではなく、様々な変形例が含まれる。例えば、上記した実施例は本発明を分かりやすく説明するために詳細に記載したものであり、必ずしも説明した全ての構成を備えるものに限定されるものではない。また、ある実施例の構成の一部を他の実施例の構成に置き換えることが可能であり、また、ある実施例の構成に他の実施例の構成を加えることも可能である。また、各実施例の構成の一部について、他の構成の追加、削除、又は置換のいずれもが、単独で、又は組み合せても適用可能である。
【０２９８】
また、上記の各構成、機能、処理部、及び処理手段等は、それらの一部又は全部を、例えば集積回路で設計する等によりハードウェアで実現してもよい。また、上記の各構成、及び機能等は、プロセッサがそれぞれの機能を実現するプログラムを解釈し、実行することによりソフトウェアで実現してもよい。各機能を実現するプログラム、テーブル、ファイル等の情報は、メモリや、ハードディスク、ＳＳＤ（Ｓｏｌｉｄ　Ｓｔａｔｅ　Ｄｒｉｖｅ）等の記録装置、又は、ＩＣカード、ＳＤカード、ＤＶＤ等の記録媒体に置くことができる。
【０２９９】
また、制御線や情報線は説明上必要と考えられるものを示しており、製品上必ずしも全ての制御線や情報線を示しているとは限らない。実際には殆ど全ての構成が相互に接続されていると考えてもよい。
【符号の説明】
【０３００】
１１　プロセッサ
１２　メモリ
１３　ストレージ装置
１０１　強化学習部
１０２　エピソード分析部
１０３　エピソードデータ学習部
１０４　オペレータ
１０５　行動制御計画装置
１１１　条件設定
１１２　シミュレータ
１１３　状態更新
１１４　行動選択
１１５　学習器
１１６　訓練データ取得

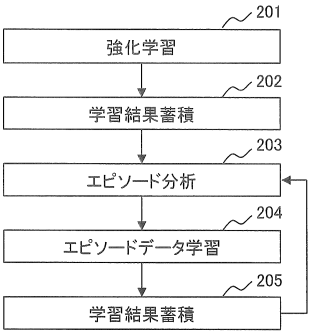
１１７　深層ニューラルネットワーク訓練
１２１　エピソード分析
１２２　エピソードデータベース
１３１　学習パターン設定
１３２　エピソード選択
１３３　エピソードデータ学習
４００　シミュレータ
４０１　行動環境設定機能
４０２　移動体行動処理機能
４０３　移動体状態量算出機能
４０４　移動体状態判定機能
４１０　強化学習／エピソードデータ学習部
４１１　方策・状態価値関数構成機能
４１２　エピソード処理機能
４１３　報酬設定機能
４１４　ＤＮＮ訓練機能
４１５　ＤＮＮ訓練用データサンプリング機能
４１６　方策・状態価値算出機能
４１７　エピソードデータ蓄積機能
４１８　行動環境データベース
４１９　エピソードデータベース
４２０　エピソードデータ学習部
４２１　学習パターン取得機能
４２２　エピソードデータ収集・一覧化機能
４２３　エピソードデータ学習機能
４３０　エピソード分析部
４３１　エピソードデータ収集機能
４３２　エピソードデータ描画機能
４３３　ステップ毎エピソード集計機能
４３４　学習パターン構成機能

【図面】

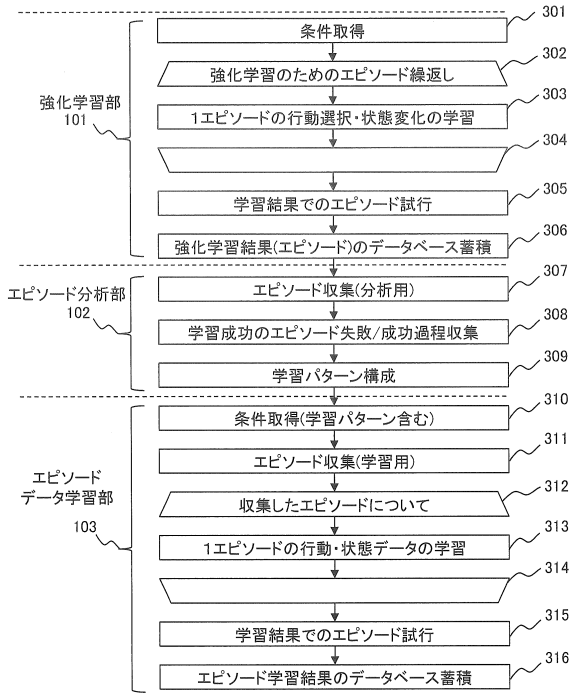
【図１】



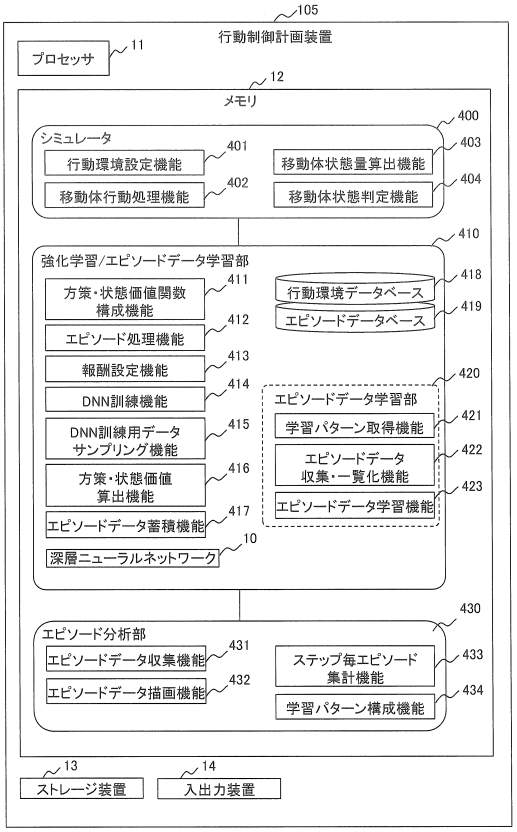
【図２】



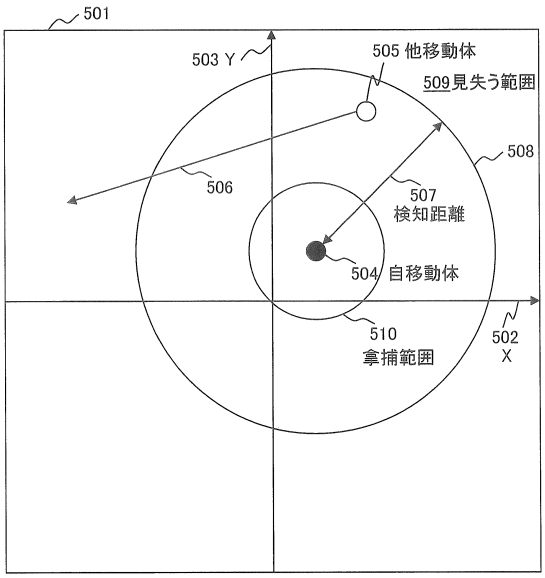
【図３】


強化学習部
101
エピソード分析部
102
エピソード
データ学習部
103
301 条件取得
302 強化学習のためのエピソード繰返し
303 1エピソードの行動選択・状態変化の学習
304
305 学習結果でのエピソード試行
306 強化学習結果(エピソード)のデータベース蓄積
307 エピソード収集(分析用)
308 学習成功のエピソード失敗/成功過程収集
309 学習パターン構成
310 条件取得(学習パターン含む)
311 エピソード収集(学習用)
312 収集したエピソードについて
313 1エピソードの行動・状態データの学習
314
315 学習結果でのエピソード試行
316 エピソード学習結果のデータベース蓄積

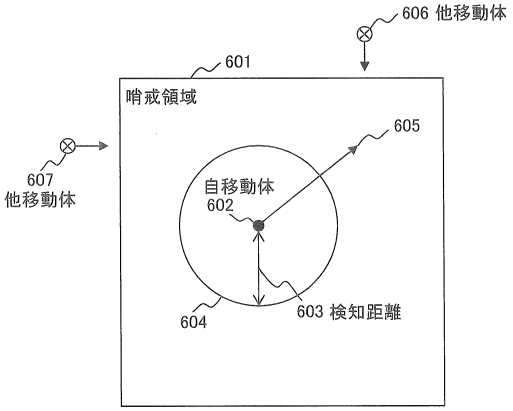
【図４】


105 行動制御計画装置
プロセッサ 11
メモリ 12
400
シミュレータ
401 行動環境設定機能
402 移動体行動処理機能
403 移動体状態量算出機能
404 移動体状態判定機能
410
強化学習/エピソードデータ学習部
411 方策・状態価値関数構成機能
412 エピソード処理機能
413 報酬設定機能
414 DNN訓練機能
415 DNN訓練用データサンプリング機能
416 方策・状態価値算出機能
417 エピソードデータ蓄積機能
10 深層ニューラルネットワーク
418 行動環境データベース
419 エピソードデータベース
420 エピソードデータ学習部
421 学習パターン取得機能
422 エピソードデータ収集・一覧化機能
423 エピソードデータ学習機能
430
エピソード分析部
431 エピソードデータ収集機能
432 エピソードデータ描画機能
433 ステップ毎エピソード集計機能
434 学習パターン構成機能
13 ストレージ装置
14 入出力装置

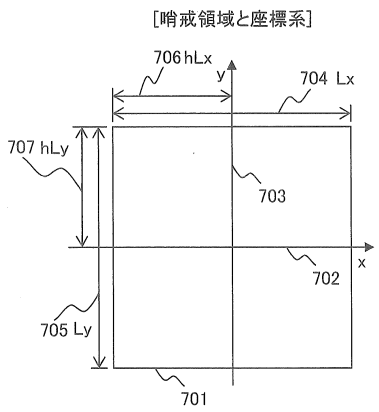
【図５】


501
503 Y
505 他移動体
509見失う範囲
508
506
507
検知距離
504 自移動体
510
拿捕範囲
502
X

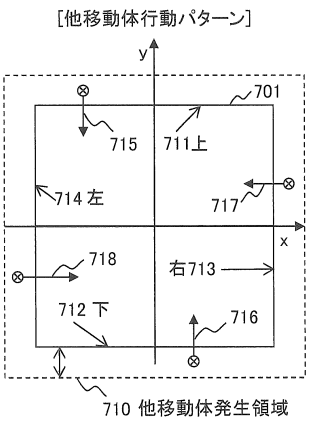
【図６】


606 他移動体
601
哨戒領域
605
607
他移動体
自移動体
602
604
603 検知距離

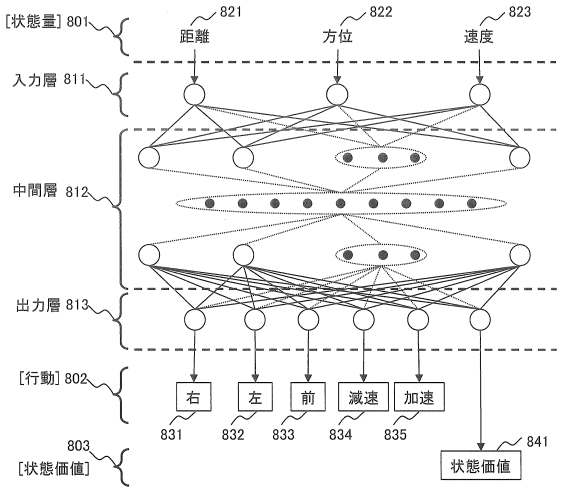
【図７Ａ】



【図７Ｂ】



【図８】



【図９Ａ】

900A ステップデータ（失敗例）

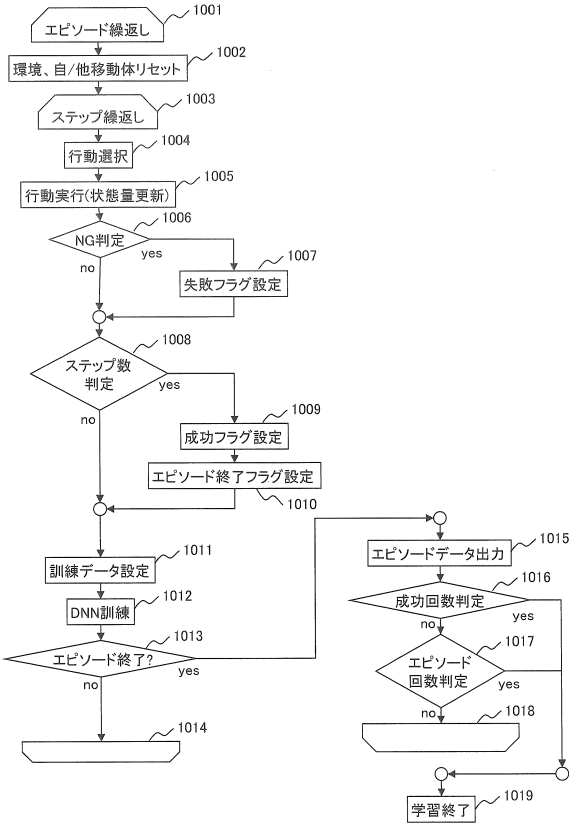
901	902			903					904	905
ステップ	状態量			行動					報酬	状態
	距離	方位	速度	前	右	左	減速	加速		
1	0.2	0.7	0.3	1	0	0	0	0	0	好位置
2	0.23	0.68	0.3	0	1	0	0	0	0	好位置
…										
273	0.5	0.12	0.1	1	0	0	0	0	−1	見失い
	911	912	913	914	915	916	917	918		

【図９Ｂ】

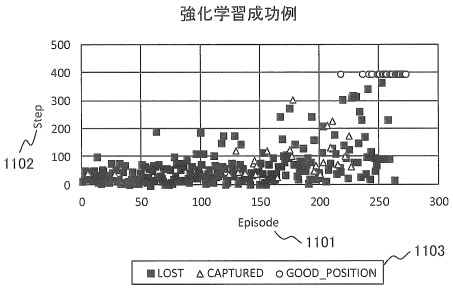
900B ステップデータ(成功例)

901 ステップ	902 状態量			903 行動					904 報酬	905 状態
	距離	方位	速度	前	右	左	減速	加速		
1	0.2	0.7	0.3	1	0	0	0	0	0	好位置
2	0.23	0.68	0.3	0	1	0	0	0	0	好位置
…										
400	0.18	0.08	0.4	0	0	0	1	0	1	好位置
	911	912	913	914	915	916	917	918		

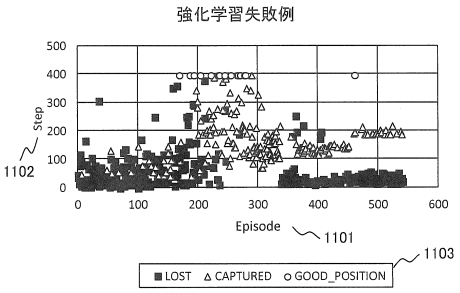
【図１０】



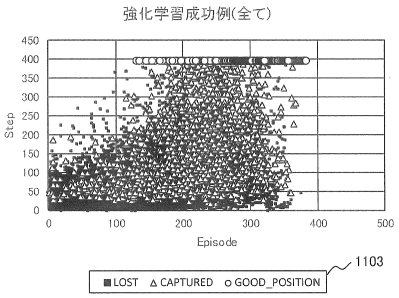
【図１１Ａ】



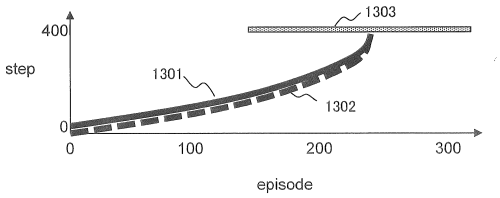
【図１１Ｂ】



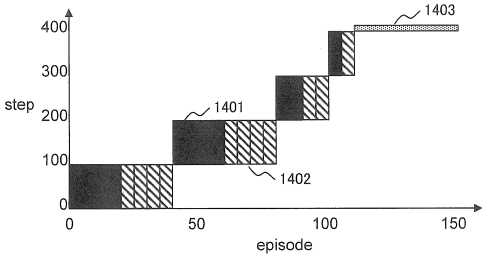
【図１２】



【図１３】



【図１４】



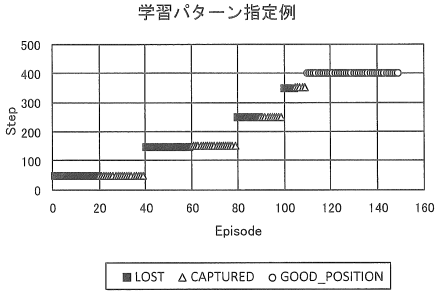
【図１５】

1500 学習パターン

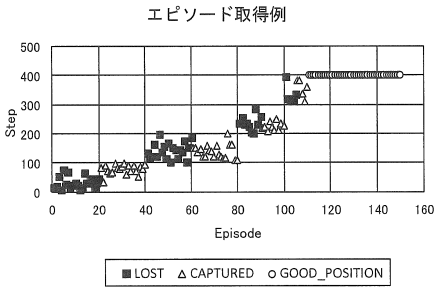
エピソード結果	終了ステップ数範囲		エピソード数
	下限	上限	
見失い	1	100	20
拿捕	1	100	20
見失い	101	200	20
拿捕	101	200	20
見失い	201	300	10
拿捕	201	300	10
見失い	301	400	5
拿捕	301	400	5
好位置	400	400	40

1501 1502 1503 1504

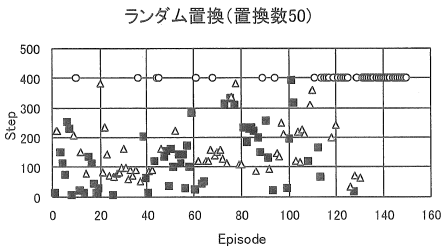
【図１６】



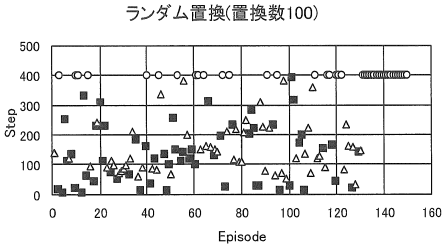
【図１７】



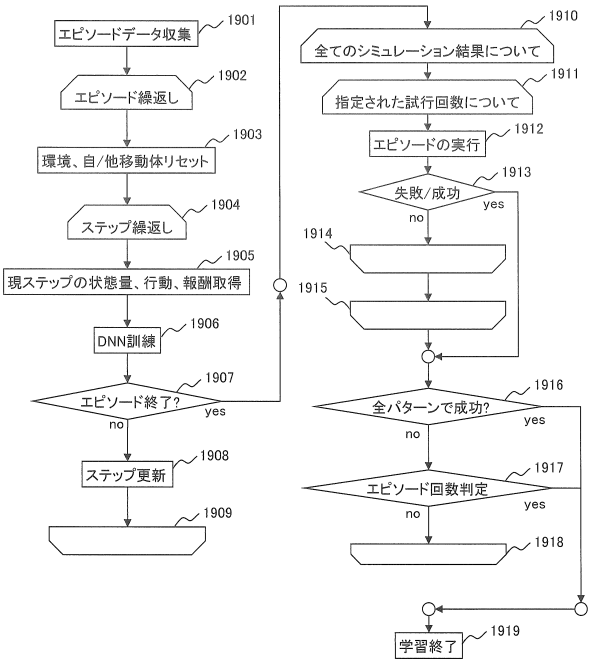
【図１８Ａ】



【図１８Ｂ】



【図１９】



フロントページの続き

(56)参考文献 特開２０２１－１８４８４（ＪＰ，Ａ）
特開２０２０－１９０８５４（ＪＰ，Ａ）
特開２０２０－１６２４３８（ＪＰ，Ａ）
特表２０１９－５２９１３５（ＪＰ，Ａ）
韓国登録特許第１０－２１６９８７６（ＫＲ，Ｂ１）
松本耕平, 外3名，“ 予測状態表現に基づく深層強化学習を用いた動的環境下における移動ロボットナビゲーション ”，第３８回日本ロボット学会学術講演会予稿集ＤＶＤ－ＲＯＭ ２０２０年，日本，一般社団法人日本ロボット学会，2020年10月09日，p.1-4

(58)調査した分野 (Int.Cl.，ＤＢ名)
Ｇ０６Ｎ ２０／００