(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2020-190930
(P2020-190930A)
(43)公開日　令和2年11月26日(2020.11.26)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 16/58	(2019.01)	G06F	16/58		5E555
G06T 7/00	(2017.01)	G06T	7/00	350C	5L096
G06F 3/0484	(2013.01)	G06F	3/0484		
G06F 16/56	(2019.01)	G06F	16/56		
G10L 15/22	(2006.01)	G10L	15/22	453	

審査請求 未請求　請求項の数 6　OL　（全 24 頁）

(21)出願番号　特願2019-95922(P2019-95922)
(22)出願日　令和1年5月22日(2019.5.22)

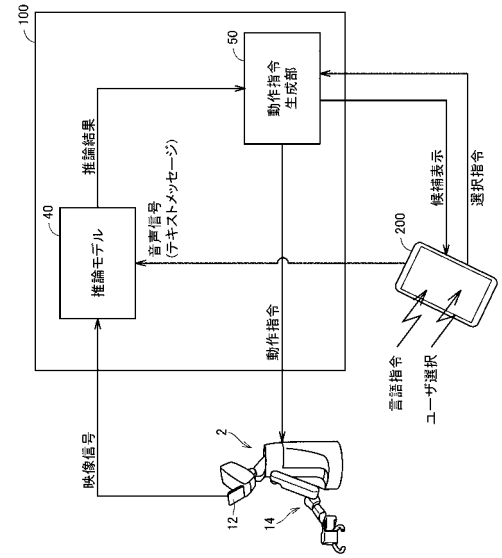
(71)出願人　301022471
国立研究開発法人情報通信研究機構
東京都小金井市貫井北町４－２－１
(74)代理人　110001195
特許業務法人深見特許事務所
(72)発明者　マガスーバ　アリー
東京都小金井市貫井北町４－２－１　国立研究開発法人情報通信研究機構内
(72)発明者　杉浦　孔明
東京都小金井市貫井北町４－２－１　国立研究開発法人情報通信研究機構内
(72)発明者　河井　恒
東京都小金井市貫井北町４－２－１　国立研究開発法人情報通信研究機構内

最終頁に続く

(54)【発明の名称】対象物検索システム、対象物検索方法および学習済モデル

(57)【要約】
【課題】対象となる物体の候補が複数存在するような状況であっても、対象となる物体を特定できる技術を提供する。
【解決手段】対象物検索システムは、特定の対象物に関する命令文を取得する命令文取得部と、命令文に関連付けられた入力画像から、当該入力画像に含まれる個々の物体を示す１または複数の第１の部分画像を抽出する画像抽出部と、命令文と、第１の部分画像の各々と、当該第１の部分画像の画像内環境を示す情報との入力を受けて、第１の部分画像の各々が命令文により特定される対象物である確率を出力する学習済モデルとを含む。学習済モデルは、画像に含まれるいずれかの物体を特定する命令文と、当該命令文により特定される物体を示す部分画像とを含むトレーニングデータセットにより学習されている。
【選択図】図１

図1



【特許請求の範囲】
【請求項１】
特定の対象物に関する命令文を取得する命令文取得部と、
前記命令文に関連付けられた入力画像から、当該入力画像に含まれる個々の物体を示す１または複数の第１の部分画像を抽出する画像抽出部と、
前記命令文と、前記第１の部分画像の各々と、当該第１の部分画像の画像内環境を示す情報との入力を受けて、前記第１の部分画像の各々が前記命令文により特定される対象物である確率を出力する学習済モデルとを備え、
前記学習済モデルは、画像に含まれるいずれかの物体を特定する命令文と、当該命令文により特定される物体を示す部分画像とを含むトレーニングデータセットにより学習されている、対象物検索システム。
【請求項２】
前記画像抽出部は、前記命令文に関連付けられた前記入力画像から、いずれかの物体が存在する区域を示す１または複数の第２の部分画像をさらに抽出するように構成されており、
前記学習済モデルは、前記第１の部分画像と前記第２の部分画像との組み合わせの各々が前記命令文により特定される対象物である確率を出力する、請求項１に記載の対象物検索システム。
【請求項３】
前記学習済モデルを規定するパラメータは、前記第１の部分画像についてのクロスエントロピー損失関数と、前記第２の部分画像についてのクロスエントロピー損失関数とを含むコスト関数に基づいて最適化される、請求項２に記載の対象物検索システム。
【請求項４】
前記学習済モデルは、
前記命令文から第１の特徴量を抽出する第１のネットワークと、
前記第１の部分画像および当該第１の部分画像の画像内環境を示す情報から第２の特徴量を抽出する第２のネットワークと、
前記第１の特徴量および前記第２の特徴量に基づいて、前記命令文により特定される対象物である確率を算出する第３のネットワークとを含む、請求項１～３のいずれか１項に記載の対象物検索システム。
【請求項５】
特定の対象物に関する命令文を取得するステップと、
前記命令文に関連付けられた入力画像から、当該入力画像に含まれる個々の物体を示す１または複数の第１の部分画像を抽出するステップと、
前記命令文と、前記第１の部分画像の各々と、当該第１の部分画像の画像内環境を示す情報とを学習済モデルに入力して、前記第１の部分画像の各々が前記命令文により特定される対象物である確率を出力するステップとを備え、
前記学習済モデルは、画像に含まれるいずれかの物体を特定する命令文と、当該命令文により特定される物体を示す部分画像とを含むトレーニングデータセットにより学習されている、対象物検索方法。
【請求項６】
対象物検索システムを構成する学習済モデルであって、
前記学習済モデルは、
特定の対象物に関する命令文と、前記命令文に関連付けられた入力画像に含まれる個々の物体を示す１または複数の第１の部分画像の各々と、当該第１の部分画像の画像内環境を示す情報との入力を受けて、前記第１の部分画像の各々が前記命令文により特定される対象物である確率を出力するものであり、
画像に含まれるいずれかの物体を特定する命令文と、当該命令文により特定される物体を示す部分画像とを含むトレーニングデータセットにより学習されている、学習済モデル。

【発明の詳細な説明】
【技術分野】
【０００１】
本技術は、命令文の言語理解および言語理解に基づく物体探索に関する。
【背景技術】
【０００２】
お年寄りや体の不自由な人の日常生活を支援するためのニーズに対して、労働力不足などの背景もあり、生活支援ロボットなどが有効な解決手段として提案されている。例えば、家庭向けサービスロボット（ＤＳＲ：domestic service robot）を標準化するような取り組みも始まっている。
【０００３】
一方で、現時点においては、生活支援ロボットは、言語での対話能力を十分に有しておらず、生活支援ロボットに命令を与えるための手段は極めて限定されている。例えば、対象物検索タスク（object retrieval task）に関して、ユーザがさまざまな言語表現を用いることは難しく、生活支援ロボットは、ある限られた言語表現の範囲内でのみ命令を理解することができるといった程度である。
【０００４】
画像および言語理解を用いて対象物を推論する技術が提案されている（非特許文献１～３）。これらの技術においては、言語情報と画像情報との間の類似性に基づいて、画像知識および言語知識を関連付けるというアプローチが採用されている。特に、非特許文献１および２は、対象物を把持するタスクに向けられており、非特許文献３は、画像内に含まれる対象物を理解するタスクに向けられている。
【先行技術文献】
【非特許文献】
【０００５】
【非特許文献１】J. Hatori et al., "Interactively Picking Real-World Objects with Unconstrained Spoken Lnguage Instructions," in IEEE ICRA, 2018, pp. 3774-3781.
【非特許文献２】M. Shridhar and D. Hsu, "Interactive visual grounding of referring expressions for human-robot interaction," in RSS, 2018.
【非特許文献３】Yu L., Tan H., Bansal M. and Berg, T. L., "A joint speaker-listener-reinforcer model for referring expressions," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, pp. 7282-7290.
【非特許文献４】J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pretraining of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.
【非特許文献５】Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al., "Google’s neural machine translation system: Bridging the gap between human and machine translation," arXiv preprint arXiv:1609.08144, 2016.
【非特許文献６】K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
【非特許文献７】A. Magassouba, K. Sugiura, and H. Kawai, "A Multimodal Classifier Generative Adversarial Network for Carry and Place Tasks From Ambiguous Language Instructions," IEEE RA-L, vol. 3, no. 4, pp. 3113-3120, Oct 2018.
【非特許文献８】K. Sugiura and H. Kawai, "Grounded Language Understanding for Manipulation Instructions Using GAN-Based Classification," IEEE ASRU, 2017.
【非特許文献９】T. Inamura, J. T. C. Tan, K. Sugiura, T. Nagai, and H. Okada, "Development of robocup@ home simulation towards long-term large scale hri," in Robot Soccer World Cup. Springer, 2013, pp. 672-680.
【発明の概要】

【発明が解決しようとする課題】
【０００６】
現実のアプリケーションにおいては、人間が発する言語による命令だけでは、認識対象の物体を一意に特定することはできず、不確実性が残ったものとなり得る。そのため、そのような不確実性にも対応できるシステムが要望されている。
【０００７】
本技術は、対象となる物体の候補が複数存在するような状況であっても、対象となる物体を特定できる技術を提供することを目的とする。
【課題を解決するための手段】
【０００８】
本発明のある局面に従う対象物検索システムは、特定の対象物に関する命令文を取得する命令文取得部と、命令文に関連付けられた入力画像から、当該入力画像に含まれる個々の物体を示す１または複数の第１の部分画像を抽出する画像抽出部と、命令文と、第１の部分画像の各々と、当該第１の部分画像の画像内環境を示す情報との入力を受けて、第１の部分画像の各々が命令文により特定される対象物である確率を出力する学習済モデルとを含む。学習済モデルは、画像に含まれるいずれかの物体を特定する命令文と、当該命令文により特定される物体を示す部分画像とを含むトレーニングデータセットにより学習されている。
【０００９】
画像抽出部は、命令文に関連付けられた入力画像から、いずれかの物体が存在する区域を示す１または複数の第２の部分画像をさらに抽出するように構成されていてもよい。学習済モデルは、第１の部分画像と第２の部分画像との組み合わせの各々が命令文により特定される対象物である確率を出力してもよい。
【００１０】
学習済モデルを規定するパラメータは、第１の部分画像についてのクロスエントロピー損失関数と、第２の部分画像についてのクロスエントロピー損失関数とを含むコスト関数に基づいて最適化されてもよい。
【００１１】
学習済モデルは、命令文から第１の特徴量を抽出する第１のネットワークと、第１の部分画像および当該第１の部分画像の画像内環境を示す情報から第２の特徴量を抽出する第２のネットワークと、第１の特徴量および第２の特徴量に基づいて、命令文により特定される対象物である確率を算出する第３のネットワークとを含んでいてもよい。
【００１２】
第３のネットワークは、第１の特徴量および第２の特徴量の入力に対する類似性を評価する識別器と、第１の特徴量と第２の特徴量との連結結果が入力される多層パーセプトロンとを含んでいてもよい。
【００１３】
第３のネットワークは、第１の特徴量および第２の特徴量が入力される、敵対的生成ネットワークを含んでいてもよい。
【００１４】
敵対的生成ネットワークは、第２の特徴量についての条件を付して学習されてもよい。
第１のネットワークは、命令文に対してサブワード埋め込み処理を行うレイヤと、サブワード埋め込み処理が行われた結果が入力されるリカレントニューラルネットワークとを含んでいてもよい。
【００１５】
対象物検索システムは、命令文により特定される対象物である確率が相対的に高い複数の第１の部分画像を出力する手段と、出力された複数の第１の部分画像に対するユーザ選択に応答して、選択された第１の部分画像に対応する物体に対して物理的な作用を与えるための動作指令を生成する手段とをさらに含む。
【００１６】

本発明の別の局面に従う対象物検索方法は、特定の対象物に関する命令文を取得するステップと、命令文に関連付けられた入力画像から、当該入力画像に含まれる個々の物体を示す１または複数の第１の部分画像を抽出するステップと、命令文と、第１の部分画像の各々と、当該第１の部分画像の画像内環境を示す情報と学習済モデルに入力して、第１の部分画像の各々が命令文により特定される対象物である確率を出力するステップとを含む。学習済モデルは、画像に含まれるいずれかの物体を特定する命令文と、当該命令文により特定される物体を示す部分画像とを含むトレーニングデータセットにより学習されている。

【００１７】

本発明のさらに別の局面に従えば、対象物検索システムを構成する学習済モデルが提供される。学習済モデルは、特定の対象物に関する命令文と、命令文に関連付けられた入力画像に含まれる個々の物体を示す１または複数の第１の部分画像の各々と、当該第１の部分画像の画像内環境を示す情報との入力を受けて、第１の部分画像の各々が命令文により特定される対象物である確率を出力するものであり、画像に含まれるいずれかの物体を特定する命令文と、当該命令文により特定される物体を示す部分画像とを含むトレーニングデータセットにより学習されている。

【発明の効果】

【００１８】

本技術によれば、対象となる物体の候補が複数存在するような状況であっても、対象となる物体を特定できる。

【図面の簡単な説明】

【００１９】

【図１】本実施の形態に従う対象物検索システムのシステム概要を示す模式図である。

【図２】本実施の形態に従う対象物検索システムの処理内容の概略を説明するための図である。

【図３】本実施の形態に従う情報処理装置のハードウェア構成例を示す模式図である。

【図４】本実施の形態に従う対象物検索システムにおいて採用される推論モデルの処理内容を説明するための図である。

【図５】本実施の形態に従う対象物検索システムにおいて採用される推論モデルの改良された処理内容を説明するための図である。

【図６】本実施の形態に従う対象物検索システムが提供する推論結果の一例を示す模式図である。

【図７】本実施の形態に従う対象物検索システムにおいて利用されるトレーニングデータセットの生成手順を示すフローチャートである。

【図８】本実施の形態に従う対象物検索システムにおいて利用される物体認識技術の結果例を示す図である。

【図９】本実施の形態に従う対象物検索システムにおけるトレーニングの処理手順を示すフローチャートである。

【図１０】本実施の形態に従う対象物検索システムにおける推論処理の処理手順を示すフローチャートである。

【発明を実施するための形態】

【００２０】

本発明の実施の形態について、図面を参照しながら詳細に説明する。なお、図中の同一または相当部分については、同一符号を付してその説明は繰り返さない。

【００２１】

［Ａ．システム概要］

まず、本実施の形態に従う対象物検索システム１の概要について説明する。図１は、本実施の形態に従う対象物検索システム１のシステム概要を示す模式図である。図１を参照して、対象物検索システム１は、典型的には、ロボット２と、情報処理装置１００と、端末装置２００とを含む。情報処理装置１００は、機能モジュールとして、推論モデル４０

および動作指令生成部５０を含む。

【００２２】

情報処理装置１００には、ロボット２に設けられたカメラ１２からの映像信号が入力される。ロボット２のカメラ１２の視野は、ユーザの視野と少なくとも一部は重複しているものとする。この状況に応じて、ユーザが端末装置２００に向けて任意の言語命令（発話による命令文）を与えると、端末装置２００を介して情報処理装置１００へ音声信号が入力される。

【００２３】

情報処理装置１００の推論モデル４０は、学習済モデルであり、端末装置２００からの音声信号と、ロボット２のカメラ１２からの映像信号との入力を受けて、推論結果を算出する。推論結果は、カメラ１２により撮像された画像内に含まれる１または複数の対象物について、ユーザからの言語命令により指定された対象物である確率を含む。

【００２４】

情報処理装置１００の動作指令生成部５０は、推論モデル４０からの推論結果を受けて、端末装置２００に操作対象の対象物の候補を表示するとともに、対象物の候補のうち、ユーザにより選択された対象物を示す選択指令を端末装置２００から受付ける。動作指令生成部５０は、選択指令に従って、対象の対象物を決定するとともに、対応する動作指令を生成して、ロボット２へ出力する。ロボット２は、動作指令に従って対象物に対する物理的な作用を与える作用部１４を駆動する。

【００２５】

このように、本実施の形態に従う対象物検索システム１においては、ユーザによる言語命令に応じて、画像内に存在する各対象物が言語命令によって指定された対象物である確率を推論する。対象物検索システム１は、このような推論結果を用いることで、ロボット２に対して、適切な動作指令を与えることができる。

【００２６】

次に、本実施の形態に従う対象物検索システム１における処理内容の概略を説明する。図２は、本実施の形態に従う対象物検索システム１の処理内容の概略を説明するための図である。図２を参照して、対象物検索システム１においては、ユーザが発した音声信号２０およびロボット２に設けられたカメラ１２などにより撮像された入力画像３０が取得される。

【００２７】

推論モデル４０には、音声信号２０に対応する命令文２２が入力される。命令文２２は、音声信号２０を公知の音声認識することでテキスト化することで生成できる。なお、音声信号２０に代えて、ユーザがキーボードといった任意の入力デバイスを用いて、命令文２２の内容を示すテキストを入力するようにしてもよい。情報処理装置１００は、命令文２２を取得する命令文取得機能として、音声認識の機能あるいはテキストベースの命令文２２を受付ける機能を有している。

【００２８】

図２には、一例として、「Bring me the toy on the wagon」（そのワゴン上のそのおもちゃを取って）といった命令文２２を示す。このように、命令文２２は、特定の対象物に関するものであるとする。

【００２９】

入力画像３０は、命令文２２が発せられたシーンを示すものである。入力画像３０からは、命令文２２により操作の対象となり得る物体を示す部分画像（以下、「ターゲット画像」あるいは「ターゲット」とも称す。）と、操作の対象となり得る物体が存在し得る区域を示す部分画像（以下、「ソース画像」あるいは「ソース」とも称す。）とが抽出される。

【００３０】

本明細書において、「命令文」は、任意の対象物に対する任意の操作を命令するものを意味する。「操作」の内容としては、例えば、対象物を「把持する」、「取る」、「しま

う」、「置く」、「移動する」といった動作が想定される。但し、これら列挙した動作に限らず、任意の操作を対象とし得る。

【００３１】

図２には、入力画像３０から複数のターゲット３２および複数のソース３４が抽出されている例を示す。推論モデル４０には、入力画像３０から抽出された１または複数のターゲット３２からなるターゲット候補３６と、入力画像３０から抽出された１または複数のソース３４からなるソース候補３８とが入力される。

【００３２】

推論モデル４０は、命令文２２、ターゲット候補３６およびソース候補３８の入力を受けて、命令文２２による操作の対象物である「たしからしさ」（likelihood）を、ターゲット３２とソース３４との各組み合わせ５２について算出する。図２に示す例では、算出される「たしからしさ」は、「そのワゴン上のそのおもちゃ」と指定されている対象物である確率を意味する。

【００３３】

このように、学習済モデルである推論モデル４０は、ターゲット３２とソース３４との組み合わせの各々が命令文２２により特定される対象物である確率を出力する。但し、ソース候補３８として１つのソース３４のみが含まれる場合には、情報処理装置１００は、ターゲット３２の各々が命令文２２により特定される対象物である確率を出力することになる。

【００３４】

さらに、ターゲット３２とソース３４との組み合わせ５２のうち、たしからしさが上位のものだけが推論結果５４として出力されてもよい。推論結果５４の内容は、端末装置２００に表示され、ユーザから最終的な選択を受付けるようにしてもよい。

【００３５】

［Ｂ．情報処理装置１００のハードウェア構成］

次に、本実施の形態に従う情報処理装置１００のハードウェア構成の一例について説明する。図３は、本実施の形態に従う情報処理装置１００のハードウェア構成例を示す模式図である。情報処理装置１００は、典型的には、汎用コンピュータを用いて実現される。

【００３６】

図３を参照して、情報処理装置１００は、主要なコンポーネントとして、プロセッサ１０２と、主メモリ１０４と、ディスプレイ１０６と、入力デバイス１０８と、ネットワークインターフェイス（Ｉ／Ｆ：interface）１１０と、光学ドライブ１１２と、入力インターフェイス（Ｉ／Ｆ）１１４と、出力インターフェイス（Ｉ／Ｆ）１１６と、二次記憶装置１２０とを含む。これらのコンポーネントは、内部バス１１８を介して互いに接続される。

【００３７】

プロセッサ１０２は、後述するような各種プログラムを実行することで、後述するような処理および機能を実現する演算主体であり、例えば、１または複数のＣＰＵ（Central Processing Unit）やＧＰＵ（Graphics Processing Unit）などで構成される。複数のコアを有するようなＣＰＵまたはＧＰＵを用いてもよい。

【００３８】

主メモリ１０４は、プロセッサ１０２がプログラムを実行するにあたって、プログラムコードやワークメモリなどを一時的に格納する記憶領域であり、例えば、ＤＲＡＭ（Dynamic Random Access Memory）やＳＲＡＭ（Static Random Access Memory）などの揮発性メモリデバイスなどで構成される。

【００３９】

ディスプレイ１０６は、処理に係るユーザインターフェイスや処理結果などを出力する表示部であり、例えば、ＬＣＤ（Liquid Crystal Display）や有機ＥＬ（Electroluminescence）ディスプレイなどで構成される。

【００４０】

入力デバイス１０８は、ユーザからの命令や操作などを受付けるデバイスであり、例えば、キーボード、マウス、タッチパネル、ペンなどで構成される。また、入力デバイス１０８としては、機械学習に必要な音声を収集するためのマイクロフォンを含んでいてもよいし、機械学習に必要な音声を収集した集音デバイスと接続するためのインターフェイスを含んでいてもよい。

【００４１】

ネットワークインターフェイス１１０は、インターネット上またはイントラネット上の任意の情報処理装置などとの間でデータを遣り取りする。ネットワークインターフェイス１１０としては、例えば、イーサネット（登録商標）、無線ＬＡＮ（Local Area Network）、Ｂｌｕｅｔｏｏｔｈ（登録商標）などの任意の通信方式を採用できる。

【００４２】

光学ドライブ１１２は、ＣＤ－ＲＯＭ（Compact Disc Read Only Memory）、ＤＶＤ（Digital Versatile Disc）などの光学ディスク１１２Ｍに格納されている情報を読出して、内部バス１１８を介して他のコンポーネントへ出力する。光学ディスク１１２Ｍは、非一過的（non-transitory）な記録媒体の一例であり、任意のプログラムを不揮発的に格納した状態で流通する。光学ドライブ１１２が光学ディスク１１２Ｍからプログラムを読み出して、二次記憶装置１２０などにインストールすることで、コンピュータにより対象物検索システム１の機能を提供できるようになる。したがって、本発明の主題は、二次記憶装置１２０などにインストールされたプログラム自体、または、本実施の形態に従う機能や処理を実現するためのプログラムを格納した光学ディスク１１２Ｍなどの記録媒体でもあり得る。

【００４３】

図３には、非一過的な記録媒体の一例として、光学ディスク１１２Ｍなどの光学記録媒体を示すが、これに限らず、フラッシュメモリなどの半導体記録媒体、ハードディスクまたはストレージテープなどの磁気記録媒体、ＭＯ（Magneto-Optical disk）などの光磁気記録媒体を用いてもよい。

【００４４】

入力インターフェイス１１４は、カメラなどの外部デバイスと接続され、カメラにより撮像された映像信号を取込む。出力インターフェイス１１６は、ロボット２などの外部デバイスと接続され、操作可能性の推論結果およびユーザからの命令などに基づいて、必要な動作指令をロボット２へ出力する。入力インターフェイス１１４および出力インターフェイス１１６は、ＵＳＢ（Universal Serial Bus）などの汎用的な通信インターフェイスを用いることができる。

【００４５】

二次記憶装置１２０は、プロセッサ１０２にて実行されるプログラム、後述するようなモデル（ニューラルネットワーク）をトレーンニングするためのトレーニングデータセット、および、モデルを規定するパラメータなどを格納するコンポーネントであり、例えば、ハードディスク、ＳＳＤ（Solid State Drive）などの不揮発性記憶装置で構成される。

【００４６】

より具体的には、二次記憶装置１２０は、図示しないＯＳ（Operating System）の他、音声認識プログラム１２１と、画像抽出プログラム１２２と、トレーニングプログラム１２３と、動作指令生成プログラム１２４と、モデルパラメータ１２５とを格納している。また、二次記憶装置１２０には、トレーニングデータセット１２６が格納されていてもよい。

【００４７】

音声認識プログラム１２１は、後述するように、音声信号２０に対応する命令文２２を生成する。画像抽出プログラム１２２は、入力画像に含まれる部分画像を抽出する（図４画像抽出部４０３に対応）。トレーニングプログラム１２３は、１２６を用いて、推論モデル４０を規定するパラメータを最適化する。動作指令生成プログラム１２４は、動作指

令生成部５０（図１）を実現する。モデルパラメータ１２５は、学習済モデルである推論モデルを規定する１または複数のパラメータを含む。トレーニングデータセット１２６は、推論モデルを最適化するための教師データであり、後述するようなデータの組からなる。

【００４８】

これらのプログラムをプロセッサ１０２で実行する際に必要となるライブラリや機能モジュールの一部を、ＯＳが標準で提供するライブラリまたは機能モジュールを用いて代替するようにしてもよい。この場合には、各プログラム単体では、対応する機能を実現するために必要なプログラムモジュールのすべてを含むものにはならないが、ＯＳの実行環境下にインストールされることで、必要な機能を実現できる。このような一部のライブラリまたは機能モジュールを含まないプログラムであっても、本発明の技術的範囲に含まれ得る。

【００４９】

また、これらのプログラムは、上述したようないずれかの記録媒体に格納されて流通するだけでなく、インターネットまたはイントラネットを介してサーバ装置などからダウンロードすることで配布されてもよい。

【００５０】

図３には、単一のコンピュータが情報処理装置１００を構成する例を示すが、これに限らず、コンピュータネットワークを介して接続された複数のコンピュータが明示的または黙示的に連携して、情報処理装置１００を含む対象物検索システム１を実現するようにしてもよい。複数のコンピュータが連携する場合、一部のコンピュータがいわゆるクラウドコンピュータと称される、ネットワーク上にある不特定のコンピュータであってもよい。

【００５１】

コンピュータ（プロセッサ１０２）がプログラムを実行することで実現される機能の全部または一部を、集積回路などのハードワイヤード回路（hard-wired circuit）を用いて実現してもよい。例えば、ＡＳＩＣ（Application Specific Integrated Circuit）やＦＰＧＡ（Field-Programmable Gate Array）などを用いて実現してもよい。

【００５２】

当業者であれば、本発明が実施される時代に応じた技術を適宜用いて、適切なハードウェア構成を採用するであろう。

【００５３】

［Ｃ．推論モデル４０（ＭＴＣＭ）］

（ｃ１：概要）

次に、本実施の形態に従う対象物検索システム１において採用される推論モデル４０の概要について説明する。

【００５４】

本実施の形態においては、推論モデル４０として、ＭＴＣＭ（Multimodal Target-source Classifier Model）と略称するモデルを採用した場合を示す。ＭＴＣＭは、上述した非特許文献１～３に開示されるような、マルチモーダル類似性ベースの統合アプローチ（multimodal similarity-based integration approach）の改良である。

【００５５】

図４は、本実施の形態に従う対象物検索システム１において採用される推論モデル４０の処理内容を説明するための図である。図４を参照して、命令文２２を処理するネットワークである命令文処理部４０１には、マルチレイヤ双方向ＬＳＴＭ（Long short-term memory）を採用する。併せて、推論精度を高める目的で、ＢＥＲＴモデル（非特許文献４などを参照）を用いて、サブワード埋め込み処理を付加している。

【００５６】

画像情報を処理するネットワークには、ＣＮＮ（Convolutional Neural Network）モデル４０４を採用している。

【００５７】

より具体的には、画像抽出部４０３が入力画像３０を処理することで、ターゲット候補３６に含まれるｉ番目（ｉ∈｛１，．．．，Ｎ｝）のターゲット３２、および、ソース候補３８に含まれるｉ’番目（ｉ’∈｛１，．．．，Ｍ｝）のソース３４の各々について、部分画像および画像内の位置が取得される。入力画像３０は、命令文２２に関連付けられたものである。そして、画像抽出部４０３は、入力画像３０に含まれる個々の物体を示す１または複数のターゲット３２（部分画像）を抽出するとともに、いずれかの物体が存在する区域を示す１または複数のソース３４（部分画像）を抽出する。

【００５８】

任意のターゲット３２について、入力データセットｘ（ｉ）を以下の（１）式のように示すことができる。

【００５９】

$$x(i)=\{x_{ins}(i),x_{v}(i),x_{rel}(i)\} \quad \cdots(1)$$

但し、x_{ins}は言語特徴量を示し、x_{v}は画像特徴量を示し、x_{rel}は関連性特徴量を示す。以下の説明においては、添え字ｉを省略して、「入力データセットｘ」とも記載する。

【００６０】

画像特徴量x_{v}は、ターゲット３２として抽出された部分画像に対応する。画像特徴量x_{v}は、ＣＮＮモデル４０４によって処理される。関連性特徴量x_{rel}は、各ターゲット３２の画像内環境（例えば、他のターゲットとの相対関係、入力画像内の位置、ソースに対する位置など）を示す情報である。

【００６１】

画像特徴量x_{v}の処理と並列的に、言語特徴量x_{ins}は、埋め込み処理がされた上で、マルチレイヤ双方向ＬＳＴＭによりエンコーディングされる。

【００６２】

言語特徴量x_{ins}および画像特徴量x_{v}を処理することで得られるそれぞれの潜在表現（latent representation）同士を比較するために、３つのＭＬＰ（多層パーセプトロン：multilayer perceptron）が配置されている。

【００６３】

最終的に、推論モデル４０からは、ターゲット３２毎の「たしからしさ」を示す推論結果が出力される。このように、推論モデル４０には、命令文２２と、ターゲット３２の各々と、ターゲット３２の画像内環境を示す情報（関連性特徴量x_{rel}）との入力を受けて、ターゲット３２の各々が命令文２２により特定される対象物である確率を出力する。

【００６４】

（ｃ２：命令文２２を処理するネットワーク）

次に、推論モデル４０の命令文２２を処理するネットワークについて説明する。図４を参照して、命令文処理部４０１には、音声信号２０に対応する命令文２２が入力され、入力された命令文２２は、リカレントニューラルネットワークの一例としてのマルチレイヤ双方向ＬＳＴＭにより処理されて、抽出された非音声特徴量ｏＩがＭＬＰ４０２に入力される。

【００６５】

推論モデル４０においては、マルチレイヤ双方向ＬＳＴＭの前段に、サブワード埋め込みモデルであるＢＥＲＴモデルが配置されている。サブワードモデルを用いて、マルチレイヤ双方向ＬＳＴＭに入力する埋め込みベクタを初期化する。すなわち、命令文２２を処理するネットワークは、命令文２２に対してサブワード埋め込み処理を行うレイヤ（典型例として、ＢＥＲＴモデル）と、記サブワード埋め込み処理が行われた結果が入力されるリカレントニューラルネットワーク（典型例として、マルチレイヤ双方向ＬＳＴＭ）とを含む。

【００６６】

ＢＥＲＴモデルは、双方向トランスフォーマに基づく言語エンコーディングモデルである。ＢＥＲＴモデルを用いることで、フレキシブル性およびロバスト性を高めることがで

きる。現在利用できるＢＥＲＴモデルは、３５億個のワードを用いてトレーニングされているため、頻出頻度の少ないワードであってもデータのスパース性は問題にはならない。

【００６７】

また、ＢＥＲＴモデルは、ワードベースのトークナイゼーション（ワードトークン）ではなく、サブワードのトークナイゼーション（サブワードトークン）を用いる。サブワードのトークナイゼーションは、ワードの一部分を用いたトークナイゼーションを意味する。例えば、以下のＴａｂｌｅ１に示すように、頻出頻度の少ないワードやミススペルされたワードに対しても、頻出頻度の高いワードを用いたトークンを生成できる。

【００６８】

【表１】

Table 1

Expression	(a)ワードトークン	(b)サブワードトークン
topright object	topright, object	top, right, object
sprayer	<UNK>	spray, er
greyis bottle	<UNK>, bottle	grey, is, bottle

【００６９】

なお、Ｔａｂｌｅ１において、＜ＵＮＫ＞は、トークンを生成できないことを意味する。なお、埋め込みモデル（ＢＥＲＴモデル）は、推論モデル４０がトレーニングされるに伴って、微調整されることになる。

【００７０】

ＢＥＲＴモデルから出力されるトークンがマルチレイヤ双方向ＬＳＴＭに入力される。マルチレイヤ双方向ＬＳＴＭは、公知技術であるので、ここでは詳細な説明は行わない。

【００７１】

さらに、マルチレイヤ双方向ＬＳＴＭからの出力は、ＭＬＰ４０２に入力される。ＭＬＰ４０２からは、入力された命令文２２の非音声特徴量ｏＩが出力される。

【００７２】

このように、推論モデル４０の命令文２２を処理するネットワークは、命令文処理部４０１およびＭＬＰ４０２を含み、命令文２２から非音声特徴量ｏＩを抽出する。

【００７３】

（ｃ３：画像情報を処理するネットワーク）

次に、推論モデル４０の画像情報を処理するネットワークについて説明する。

【００７４】

ＣＮＮモデル４０４としては、例えば、非特許文献６に示されるような１６層ネットワーク（ＶＧＧ１６）を用いて、画像特徴量をエンコーディングできる。ＣＮＮモデル４０４からの出力は、連結部４０５において関連性特徴量x_{rel}と連結される。

【００７５】

ターゲット候補３６に含まれるＮ個のターゲット３２の各々と、対応する関連性特徴量x_{rel}とについて、連結部４０５による連結結果が出力される。そして、すべてのターゲット３２についての連結結果がＭＬＰ４０６に入力される。ＭＬＰ４０６からは、ターゲット候補３６に含まれる複数のターゲット３２についての画像特徴量ｏＶが出力される。

【００７６】

このように、推論モデル４０の画像情報を処理するネットワークは、ターゲット３２およびターゲット３２の画像内環境を示す情報である関連性特徴量x_{rel}から画像特徴量ｏＶを抽出する。

【００７７】

（ｃ４：推論モデル４０の推論結果を生成する出力部４１０）

推論モデル４０の推論結果Ｙは、以下の（２）式のように示すことができる。

【００７８】

Ｙ＝｛ｙｔａｒｇ，ｙｓｒｃ｝　・・・（２）

但し、ｙｔａｒｇはターゲットについての推論結果を示し、ｙｓｒｃはソースについての推論結果を示す。推論結果ｙｔａｒｇおよびｙｓｒｃは、いずれもＮ×Ｍ次元のベクトルとして規定される。

【００７９】

ターゲットについての推論結果ｙｔａｒｇは、入力された命令文２２の非音声特徴量ｏＩと、ターゲット候補３６に含まれる複数のターゲット３２についての画像特徴量ｏＶとが類似性識別器４０７に入力されることで算出される。

【００８０】

ソースについての推論結果ｙｓｒｃは、入力された命令文２２の非音声特徴量ｏＩと、ターゲット候補３６に含まれる複数のターゲット３２についての画像特徴量ｏＶとが連結部４０８において連結された結果がＭＬＰ４０９に入力されることで算出される。

【００８１】

このように、出力部４１０は、非音声特徴量ｏＩおよび画像特徴量ｏＶに基づいて、各ターゲット３２が命令文２２により特定される対象物である確率を算出するネットワークである。より具体的には、このネットワークは、非音声特徴量ｏＩおよび画像特徴量ｏＶの入力に対する類似性を評価する類似性識別器４０７と、非音声特徴量ｏＩと画像特徴量ｏＶとの連結結果が入力されるＭＬＰ４０８とを含む。

【００８２】

（ｃ５：推論モデル４０のトレーニング）

推論モデル４０のコスト関数ＪＭＴＣＭは、以下の（３）式のように定義できる。

【００８３】

ＪＭＴＣＭ＝λ１Ｊｔａｒｇ＋λ２Ｊｓｒｃ　・・・（３）

但し、λ１およびλ２は重みパラメータであり、Ｊｔａｒｇは、ターゲット３２についてのクロスエントロピー損失関数であり、Ｊｓｒｃは、ソース３４についてのクロスエントロピー損失関数である。クロスエントロピー損失関数ＪｔａｒｇおよびＪｓｒｃは、以下の（４－１）および（４－２）式のように定義できる。

【００８４】

【数１】

$$J_{targ} = -\sum_{n}\sum_{m} y^{*}_{targ\_nm} \log p(y_{targ\_nm}) \quad \cdots(4\text{-}1)$$

$$J_{src} = -\sum_{n}\sum_{m} y^{*}_{src\_nm} \log p(y_{src\_nm}) \quad \cdots(4\text{-}2)$$

【００８５】

但し、ｙ*ｔａｒｇ＿ｎｍおよびｙ*ｓｒｃ＿ｎｍは、ｎ番目のサンプルのｍ番目の次元についてのラベル（正解）を示し、ｙｔａｒｇ＿ｎｍおよびｙｓｒｃ＿ｎｍは、ｎ番目のサンプルのｍ番目の次元についての推論結果を示す。

【００８６】

図４に示される推論モデル４０は、予め用意されたトレーニングデータセット１２６に対して、上述した（３）式で定義されるコスト関数ＪＭＴＣＭが最小になるようにパラメータを最適化することで構成される。このように、推論モデル４０を規定するパラメータは、ターゲット３２についてのクロスエントロピー損失関数Ｊｔａｒｇと、ソース３４についてのクロスエントロピー損失関数Ｊｓｒｃとを含むコスト関数に基づいて最適化されることになる。

【００８７】

［Ｄ．推論モデル４０（ＭＴＣＭ－ＧＡＮ）］

（ｄ１：概略）

上述したＭＴＣＭからなる推論モデル４０に対して、敵対的生成ネットワーク（ＧＡＮ：generative adversarial nets）を付加することで、トレーニングデータを増大させて

、識別性能を高めることもできる。以下、ＭＴＣＭおよびＧＡＮからなる推論モデル４０（以下、「ＭＴＣＭ－ＧＡＮ」とも記載する。）について説明する。

【００８８】

先に、ＧＡＮについて概略する。ＧＡＮフレームワークは、生成器（generator）Ｇおよび識別器（discriminator）Ｄの２つの敵対的ネットワークで構成される。生成器Ｇは、所与の分布データを模倣することで疑似データを生成する。並行して、識別器Ｄは、入力データが真（real）であるか偽（fake）であるかを識別（推論）する。これらのネットワークの目的として、生成器Ｇはより真に近いデータを生成するようになり、一方で、識別器Ｄはその識別能力（推論能力）を向上させる。

【００８９】

生成器Ｇは、正規分布からランダムにサンプルされた複数次元の入力ｚ（ノイズ）を用いて、疑似サンプルx_{fake}を生成する。真正サンプルx_{real}と疑似サンプルx_{fake}とを識別するために、ソースフラグＳ∈｛ｒｅａｌ，ｆａｋｅ｝に応じて、入力ｘ＝x_{real}または$x=x_{fake}$が識別器Ｄには選択的に入力される。識別器Ｄの出力は、推論確率ｐＤ（Ｓ＝ｒｅａｌ｜ｘ）＝Ｄ（ｘ）となる。生成器Ｇの損失関数ＪＧおよび識別器Ｄの損失関数ＪＤは、以下の（５－１）式で定義される損失関数ＪＳを用いて、（５－２）および（５－３）式のように定義できる。これらの損失関数を用いて、ＧＡＮのパラメータが最適化される。

【００９０】

【数２】

$$J_S=-\frac{1}{2}\boldsymbol{E}_{x_{real}}\log D(\boldsymbol{x}_{real})-\frac{1}{2}\boldsymbol{E}_{z}\log(1-D(\boldsymbol{x}_{fake})) \quad \cdots(5\text{-}1)$$

$$J_D=J_S \quad \cdots(5\text{-}2)$$

$$J_G=-J_S \quad \cdots(5\text{-}3)$$

【００９１】

（ｄ２：ＭＴＣＭ－ＧＡＮ）

ＧＡＮのデータ増大特性を利用して、推論モデル４０にデータ増大および識別を同時に行う機能を付加した改良例について説明する（非特許文献７および非特許文献８など参照）。

【００９２】

図５は、本実施の形態に従う対象物検索システム１において採用される推論モデル４０の改良された処理内容を説明するための図である。図５を参照して、ＭＴＣＭ－ＧＡＮは、図４に示す推論モデル４０の推論結果を生成する出力部４１０に代えて、出力部４２０を有している。推論モデル４０の出力部４２０以外の部分は、図４と同様であるので、詳細な説明は繰り返さない。

【００９３】

出力部４２０は、連結部４２１，４２２と、生成器４２３と、選択部４２４と、識別器４２５とを含む。生成器４２３および識別器４２５が敵対的ネットワークを構成する。出力部４２０においては、生成器４２３により生成される疑似データは識別器４２５の識別能力を向上させる。識別器４２５は、真正サンプルx_{real}と疑似サンプルx_{fake}とを識別するだけではなく、候補となるターゲットの「たしからしさ」を推論することで識別タスクも実行することになる。

【００９４】

そのため、出力部４２０の識別器４２５は、推論確率ｐＤ（Ｓ）に加えて、第２の出力として、各ターゲットが命令文２２による操作の対象物である確率を示す推論確率ｐＤ（y_{targ}）を出力する。

【００９５】

また、識別器４２５のコスト関数ＪＤは、以下の（６）式のように定義できる。

ＪＤ＝ＪＳ＋λＪ　・・・（６）

但し、λは重みパラメータであり、Ｊは上述した（５－１）式において定義したクロスエントロピー損失関数である。

【００９６】

ＭＴＣＭの推論モデル４０の初期状態を考慮すると、図５に示される出力部４２０の識別器４２５へ入力されるデータセットは、以下の（７）式のように設定できる。

【００９７】

ｘＧＡＮ＝｛x_{real}＝（ｏＶ，ｏＩ），x_{fake}＝（ｚ，ｏＶ）｝　・・・（７）

このように、出力部４２０は、非音声特徴量ｏＩおよび画像特徴量ｏＶが入力される、敵対的生成ネットワークを含む。敵対的生成ネットワークを用いることで、識別性能を高めることもできる。

【００９８】

（ｄ３：条件付きＭＴＣＭ－ＧＡＮ）

図５に示すＭＴＣＭ－ＧＡＮにおいて、生成器４２３に対して画像特徴量ｏＶについての条件を付してもよい。これは、生成器４２３および識別器４２５が全結合ネットワークであるとともに、画像特徴量ｏＶが、生成器４２３（真正サンプルx_{real}を通じて）および識別器４２５の両方に入力されるからである。すなわち、敵対的生成ネットワーク（出力部４２０）は、画像特徴量ｏＶについての条件を付して学習されてもよい。

【００９９】

より具体的には、識別器４２５は、入力ソースがいずれであるかを推論するのに加えて、組み合わせ（ｏＩ，ｏＶ）が正しいものであるか否かを推論する。一方、生成器４２３は、正解／不正解の組み合わせ（ｏＩ，ｏＶ）を疑似データとして生成する。そこで、同一のシーンのランダムに選択された不正解のターゲットｊを考慮しつつ、各ターゲットｉについて、正解の特徴量の組み合わせ（ｏＩ（ｉ），ｏＶ（ｉ））および不正解の特徴量の組み合わせ（ｏＩ（ｉ），ｏＶ（ｊ））を用意し、これを用いてトレーニングを行うようにしてもよい。

【０１００】

［Ｅ．推論結果の出力例］

次に、本実施の形態に従う対象物検索システム１による推論結果の出力例について説明する。本実施の形態に従う対象物検索システム１は、ターゲット３２とソース３４との組み合わせの各々について、命令文２２による操作の対象物である「たしからしさ」（すなわち、確率）を算出できる。

【０１０１】

ターゲット３２とソース３４との組み合わせについての確率、あるいは、ターゲット３２についての確率をユーザに提示するようにしてもよい。

【０１０２】

図６は、本実施の形態に従う対象物検索システム１が提供する推論結果の一例を示す模式図である。図６を参照して、対象物検索システム１は、端末装置２００のスクリーン上などに推論結果を含む結果表示３００を提供することができる。結果表示３００は、シーン（入力画像３０）から抽出されたターゲット３２の各々について、命令文２２による操作の対象物である確率６０が表示されている。

【０１０３】

図６に示すように、結果表示３００にはターゲット３２毎の確率６０が表示されているので、ユーザは、意図したターゲット３２をより容易に選択できる。また、算出される確率を定量的に評価できるので、しきい値などの条件に基づいて、対象となるターゲット３２を自動的に選択することができる。

【０１０４】

［Ｆ．処理手順］

次に、本実施の形態に従う対象物検索システム１における処理手順について説明する。

【０１０５】
（ｆ１：トレーニングデータセット１２６の生成）
図７は、本実施の形態に従う対象物検索システム１において利用されるトレーニングデータセット１２６の生成手順を示すフローチャートである。図７に示す各ステップは、コンピュータにより実行されてもよいし、一部をユーザ自身が実行してもよい。
【０１０６】
図７を参照して、ユーザは、シーンを示す１または複数の入力画像を取得する（ステップＳ１００）。入力画像は、現実の室内を撮像することで取得してもよいし、画像共有サイトなどから任意にダウンロードすることで取得してもよい。
【０１０７】
続いて、取得された１または複数の入力画像のうち１つを選択し（ステップＳ１０２）、選択された入力画像から物体を示す１または複数の領域（ターゲット３２）を抽出する（ステップＳ１０４）。抽出された１または複数のターゲット３２のうち１つを選択し（ステップＳ１０６）、選択されたターゲット３２が存在する区域を示す部分画像（ソース３４）を抽出する（ステップＳ１０８）。さらに、選択されたターゲット３２に対して、対応する物体の名称をラベルとして付与する（ステップＳ１１０）。併せて、選択されたターゲット３２とシーンとの関連性を示す情報（関連性特徴量x_{rel}）を設定する（ステップＳ１１２）。
【０１０８】
ステップＳ１０４およびＳ１１０に関して、入力画像に対して公知の物体認識技術を用いて自動的にターゲット３２となり得る領域を抽出するようにしてもよい。
【０１０９】
図８は、本実施の形態に従う対象物検索システム１において利用される物体認識技術の結果例を示す図である。図８に示すように、入力された画像に対して、物体が存在する領域が特定および抽出されるとともに、特定された物体を示すラベル（例えば、物品名）が自動的に抽出される。このような抽出結果を用いて、ターゲット３２および対応するラベルのデータセットを自動的に生成できる。
【０１１０】
このような物体認識技術としては、ＳＳＤ（Single Shot MultiBox Detector）やＹＯＬＯ（You Only Look Once）などのアルゴリズムを用いることができる。
【０１１１】
あるいは、領域抽出およびラベル付与を手動で行うようにしてもよい。さらにあるいは、公知のアルゴリズムを用いて自動的に領域を抽出した上で、手動でラベルを付与するようにしてもよい。
【０１１２】
また、ソース３４については、ターゲット３２を囲むような領域を抽出するようにしてもよい。
【０１１３】
ステップＳ１１２に関して、抽出したターゲット３２とシーンとの関連性を示す情報としては、例えば、「右下」や「左上」といった自然言語表現であってもよいし、位置の情報を示す符号であってもよい。
【０１１４】
さらに、選択されたターゲット３２に関する１または複数の命令文２２を取得する（ステップＳ１１４）。１または複数の命令文２２は、ユーザが任意に考えて設定してもよい。例えば、いわゆるクラウドワーカに対して入力画像を提供するとともに、対応する１または複数の命令文２２を応答してもらうような形態が想定できる。
【０１１５】
ステップＳ１０６～Ｓ１１４の処理によって、１つの入力画像に含まれる１つのターゲット３２に対応付けられる、ラベル、ソース３４、関連性特徴量x_{rel}、命令文２２からなるデータセットを取得できる。

【０１１６】
選択された入力画像に含まれるターゲット３２のすべてについて処理が完了したか否かが判断される（ステップＳ１１６）。選択された入力画像に含まれるターゲット３２のうち処理が完了していないものがあれば（ステップＳ１１６においてＮＯ）、新たなターゲット３２が選択され（ステップＳ１１８）、ステップＳ１０８以下の処理が繰り返される。
【０１１７】
選択された入力画像に含まれるすべてのターゲット３２について処理が完了していれば（ステップＳ１１６においてＹＥＳ）、取得された入力画像のすべてについて処理が完了したか否かが判断される（ステップＳ１２０）。取得された入力画像のうち処理が完了していないものがあれば（ステップＳ１２０においてＮＯ）、新たな入力画像が選択され（ステップＳ１２２）、ステップＳ１０４以下の処理が繰り返される。
【０１１８】
取得されたすべての入力画像について処理が完了していれば（ステップＳ１２０においてＹＥＳ）、ステップＳ１０６～Ｓ１１４の処理によって得られるデータセットがトレーニングデータセット１２６として出力される（ステップＳ１２４）。そして、処理は終了する。
【０１１９】
図７に示すトレーニングデータセット１２６の生成手順によれば、各入力画像に含まれる各ターゲット３２に対応付けられる、ラベル、ソース３４、関連性特徴量x_{rel}、命令文２２からなるデータセットを取得できる。各トレーニングデータセット１２６には、ターゲット３２およびソース３４の位置および大きさの情報を含めるようにしてもよい。
【０１２０】
（ｆ２：推論モデル４０のトレーニング）
図９は、本実施の形態に従う対象物検索システム１におけるトレーニングの処理手順を示すフローチャートである。図９に示す各ステップは、情報処理装置１００のプロセッサ１０２がトレーニングプログラム１２３を実行することで実現されてもよい。
【０１２１】
図９を参照して、情報処理装置１００は、予め用意されたトレーニングデータセット１２６のうち１つのデータセットを選択し（ステップＳ２００）、選択されたデータセットから入力データセットｘおよび対応する正解ラベル（y^*_{targ}およびy^*_{src}など）を生成する（ステップＳ２０２）。なお、ＭＴＣＭ－ＧＡＮからなる推論モデル４０を採用する場合には、正解ラベルとして、非音声特徴量ｏＩおよび画像特徴量x_vを用いて、推論確率ｐＤ（Ｓ）および推論確率ｐＤ（y_{targ}）を算出してもよい。
【０１２２】
情報処理装置１００は、予め用意されたトレーニングデータセット１２６のすべてについての処理が完了したか否かを判断する（ステップＳ２０４）。予め用意されたトレーニングデータセット１２６のうち処理が完了していないものがあれば（ステップＳ２０４においてＮＯ）、新たなデータセットが選択され（ステップＳ２０６）、ステップＳ２０２以下の処理が繰り返される。
【０１２３】
予め用意されたトレーニングデータセット１２６に含まれるすべてのデータセットについて処理が完了していれば（ステップＳ２０４においてＹＥＳ）、情報処理装置１００は、生成された入力データセットｘを推論モデル４０に入力するとともに、算出される推論結果と対応する正解ラベルとの誤差に基づいて、推論モデル４０のモデルパラメータを最適化する（ステップＳ２０８）。すなわち、学習済モデルである推論モデル４０は、入力画像に含まれるいずれかの物体を特定する命令文２２と、命令文２２により特定されるターゲット３２（物体を示す部分画像）とを含むトレーニングデータセット１２６により学習されることで、生成される。
【０１２４】

より具体的には、推論モデル４０のモデルパラメータの最適化には、上述したようなクロスエントロピー損失関数が用いられる。

【０１２５】

なお、バッチノーマリゼーションやドロップアウトなどの公知の加速化手法を採用できる。

【０１２６】

（ｆ３：推論モデル４０を用いた推論処理）

図１０は、本実施の形態に従う対象物検索システム１における推論処理の処理手順を示すフローチャートである。図１０に示す各ステップは、情報処理装置１００のプロセッサ１０２が各種プログラムを実行することで実現されてもよい。

【０１２７】

図１０を参照して、情報処理装置１００は、端末装置２００から音声信号２０が入力されると（ステップＳ３００）、入力された音声信号２０を音声認識してテキストベースの命令文２２を取得する（ステップＳ３０２）。このように、情報処理装置１００は、特定の対象物に関する命令文２２を取得する。

【０１２８】

並行して、情報処理装置１００は、命令文２２に関連付けられた入力画像３０から、入力画像３０に含まれる個々の物体を示す１または複数のターゲット３２（部分画像）を抽出するとともに、いずれかの物体が存在する区域を示す１または複数のソース３４（部分画像）を抽出する。より具体的には、情報処理装置１００は、入力画像３０を取得し（ステップＳ３０４）、取得した入力画像３０から１または複数のターゲット３２および１または複数のソース３４を抽出する（ステップＳ３０６）。

【０１２９】

情報処理装置１００は、ステップＳ３０６において抽出したいずれかのターゲット３２といずれかのソース３４との組み合わせを選択する（ステップＳ３０８）とともに、選択した組み合わせにおける関連性特徴量x_{rel}を決定する（ステップＳ３１０）。

【０１３０】

そして、情報処理装置１００は、命令文２２と、ターゲット３２の各々と、ターゲット３２の画像内環境を示す情報とを学習済モデルである推論モデル４０に入力して、ターゲット３２の各々が命令文２２により特定される対象物である確率を出力する。すなわち、情報処理装置１００は、情報処理装置１００は、命令文２２（ステップＳ３０２）、選択した組み合わせを構成するターゲット３２（ステップＳ３０８）および選択した組み合わせにおける関連性特徴量x_{rel}（ステップＳ３１０）を推論モデル４０に入力し、選択した組み合わせについての推論結果を算出する（ステップＳ３１２）。

【０１３１】

情報処理装置１００は、ステップＳ３０６において抽出したターゲット３２とソース３４とのすべての組み合わせについて推論結果の算出が完了したか否かを判断する（ステップＳ３１４）。推論結果の算出が完了していない組み合わせが存在していれば（ステップＳ３１４においてＮＯ）、情報処理装置１００は、ターゲット３２とソース３４との新たな組み合わせを選択し（ステップＳ３１６）、ステップＳ３１０以下の処理を実行する。

【０１３２】

すべての組み合わせについて推論結果の算出が完了していれば（ステップＳ３１４においてＹＥＳ）、情報処理装置１００は、推論結果のスコアが上位の１または複数の組み合わせを選択して、選択された組み合わせを含む画面を端末装置２００に表示する（ステップＳ３１８）。このように、情報処理装置１００は、命令文２２により特定される対象物である確率が相対的に高い複数のターゲット３２（および、対応するソース３４）を出力する。

【０１３３】

情報処理装置１００は、端末装置２００からユーザの選択指令を受けると（ステップＳ３２０）、選択指令により指定されたターゲット３２を対象として動作指令を生成する（

ステップＳ３２２）。このように、情報処理装置１００は、出力された複数のターゲット３２（および、対応するソース３４）に対するユーザ選択に応答して、選択されたターゲット３２に対応する物体に対して物理的な作用を与えるための動作指令を生成する。なお、生成された動作指令は、ロボット２などへ出力されてもよい。

【０１３４】

［Ｇ．評価結果］

次に、本実施の形態に従う対象物検索システム１の性能を評価した結果例を示す。

【０１３５】

（ｇ１：ＰＦＮ－ＰＩＣ）

まず、ＰＦＮ－ＰＩＣデータセット（非特許文献１）を用いて、本実施の形態に従う対象物検索システム１の性能を評価した。上述したように、推論モデル４０としては、ＭＴＣＭ、ＭＴＣＭに敵対的生成ネットワーク（ＧＡＮ）を付加したＭＴＣＭ－ＧＡＮ、および、ＭＴＣＭ－ＧＡＮに対して条件を付したモデルの３種類を採用可能である。それぞれのモデルについて、非特許文献１に開示される手法との比較を含めて評価を行った。

【０１３６】

より具体的には、ＰＦＮ－ＰＩＣデータセットのうち８９８６１組のデータセットを用いて推論モデル４０をトレーニングするとともに、別の８９８組のデータセットを用いて評価を行った。

【０１３７】

なお、ＰＦＮ－ＰＩＣデータセットは、ピックアンドプレイスのタスクに向けられたものであり、上から見て４つのボックスのいずれかに配置された対象物を見つけることが想定されている。各ボックスをソース３４と見なし、対象物をターゲット３２と見なして評価を行った。

【０１３８】

命令文２２としては、例えば、"Grab the black mug and put it in the lower right box."（黒いマグをつかんで右下のボックスに入れなさい。）といった、対象となるターゲット３２およびソース３４を含むものを用いた。

【０１３９】

本実施の形態に従う対象物検索システム１においては、ターゲット３２とソース３４との組み合わせ毎、あるいは、ターゲット３２毎に、命令文２２による操作の対象物である「たしからしさ」（確率）を算出できる。そのため、複数の物体が命令文２２による操作の対象物になり得る（すなわち、命令文２２が不確実性を含んでいる）場合であっても、算出される確率に基づいて処理が可能である。

【０１４０】

これに対して、非特許文献１に開示される手法は、命令文２２による操作の対象物は１つであることが前提となっており、複数の物体が対象物になり得ることは何ら想定されていない。

【０１４１】

非特許文献１は、マルチモーダル類似性ベースの手法であり、類似性の正確性（類似性についての正答率）を評価した。この類似性の正確性は、推論モデル４０としてＭＴＣＭを採用した場合に、類似性識別器４０７から出力されるターゲットについての推論結果y_{targ}（図４参照）に相当する。Ｔａｂｌｅ１においては、類似性の正確性に関しては、非特許文献１の手法とＭＴＣＭとを比較している。

【０１４２】

本実施の形態に従う対象物検索システム１においては、与えられた命令文２２の対象となるターゲット３２以外に１または複数の不正解のターゲット３２を用意した場合の正確性を評価した（領域毎の正確性）。パラメータγは、不正解のターゲット３２の数を示す。

【０１４３】

【表２】
Table 2

手法	類似性の正確性	領域毎の正確性				
		γ=1	γ=2	γ=4	γ=45	γ=10
非特許文献1	88	-	-	-	-	
MTCM	88.8	94.5	95.4	96.1	96.3	97.1
MTCM-GAN（条件無）	-	95.7	96.1	96.2	96.4	96.8
MTCM-GAN（条件付）	-	95.9	96.3	96.5	96.5	97.2

【０１４４】
Ｔａｂｌｅ２に示すように、推論モデル４０としてＭＴＣＭを採用した場合であっても、非特許文献１に開示される手法に比較して改善効果を見ることができる。さらに、条件付きＭＴＣＭ－ＧＡＮを推論モデル４０として採用することで、最も高い識別性能が発揮されていることが分かる。
【０１４５】
（ｇ２：ＷＲＳ－ＶＳ）
次に、Ｗｏｒｌｄ Ｒｏｂｏｔ Ｓｕｍｍｉｔ ２０１８ Ｖｉｒｔｕａｌ Ｓｐａｃｅ（以下、「ＷＲＳ－ＶＳ」とも称す。）ｃｈａｌｌｅｎｇｅで利用されたデータセットを用いても、本実施の形態に従う対象物検索システム１の性能を評価した。ＷＲＳ－ＶＳで用いられたデータセットは、ＳＩＧＶｅｒｓｅに基づくものである（非特許文献９など参照）。
【０１４６】
より具体的には、ＷＲＳ－ＶＳデータセットのうち１０１０組のデータセットを用いて推論モデル４０をトレーニングするとともに、別の３７組のデータセットを用いて評価を行った結果をＴａｂｌｅ３に示す。なお、パラメータγは、不正解のターゲット３２の数を示す。
【０１４７】
【表３】
Table 3

手法	領域毎の正確性	
	γ=1	γ=2
MTCM	83.5	85.6
MTCM-GAN（条件無）	89.6	92.6
MTCM-GAN（条件付）	90.7	93.6

【０１４８】
Ｔａｂｌｅ３に示すように、推論モデル４０として、条件付きＭＴＣＭ－ＧＡＮを推論モデル４０として採用することで、最も高い識別性能が発揮されていることが分かる。
【０１４９】
［Ｈ．変形例］
本実施の形態に従う対象物検索システム１（推論モデル４０）を十分に大きいトレーニングデータセットを用いてトレーニングすることで、汎化性能を高めることができる。こ

の場合、トレーニングによって得られたモデルパラメータのみを配布するようにしてもよい。

【０１５０】

上述の説明においては、タスクが実行される場所にシステムを配置する、いわゆるオンプレミス環境に適した処理例を示すが、これに限らず、コンピュータネットワーク上に配置された１または複数のサーバを用いて、タスクを処理する、いわゆるクラウドサービス環境を採用してもよい。

【０１５１】

本実施の形態に従う推論モデル４０は、要求されるタスクの内容や実行環境などに応じて適宜適切な実装が可能である。例えば、推論モデル４０を別のモデルの一部として組み込む、あるいは、推論モデル４０と別のモデルとを組み合わせるといった実装形態が可能である。

【０１５２】

［Ｉ．まとめ］

人間が発する言語による命令だけでは、認識対象の物体を一意に特定することはできず、不確実性が残ったものとなり得えるが、本実施の形態に従えば、対象となる物体の候補が複数存在するような状況であっても、対象となる物体である確率を評価しつつ、対象となる物体を容易に特定できる。

【０１５３】

今回開示された実施の形態は、すべての点で例示であって制限的なものではないと考えられるべきである。本発明の範囲は、上記した実施の形態の説明ではなくて特許請求の範囲によって示され、特許請求の範囲と均等の意味および範囲内でのすべての変更が含まれることが意図される。

【符号の説明】

【０１５４】

１　対象物検索システム、２　ロボット、１２　カメラ、１４　作用部、２０　音声信号、２２　命令文、３０　入力画像、３２　ターゲット、３４　ソース、３６　ターゲット候補、３８　ソース候補、４０　推論モデル、５０　動作指令生成部、５２　組み合わせ、５４　推論結果、６０　確率、１００　情報処理装置、１０２　プロセッサ、１０４　主メモリ、１０６　ディスプレイ、１０８　入力デバイス、１１０　ネットワークインターフェイス、１１２　光学ドライブ、１１２Ｍ　光学ディスク、１１４　入力インターフェイス、１１６　出力インターフェイス、１１８　内部バス、１２０　二次記憶装置、１２１　音声認識プログラム、１２２　画像抽出プログラム、１２３　トレーニングプログラム、１２４　動作指令生成プログラム、１２５　モデルパラメータ、１２６　トレーニングデータセット、２００　端末装置、３００　結果表示、４０１　命令文処理部、４０２，４０６，４０９　ＭＬＰ、４０３　画像抽出部、４０４　モデル、４０５，４０８，４２１，４２２　連結部、４０７　類似性識別器、４１０，４２０　出力部、４２３　生成器、４２４　選択部、４２５　識別器。

【図１】

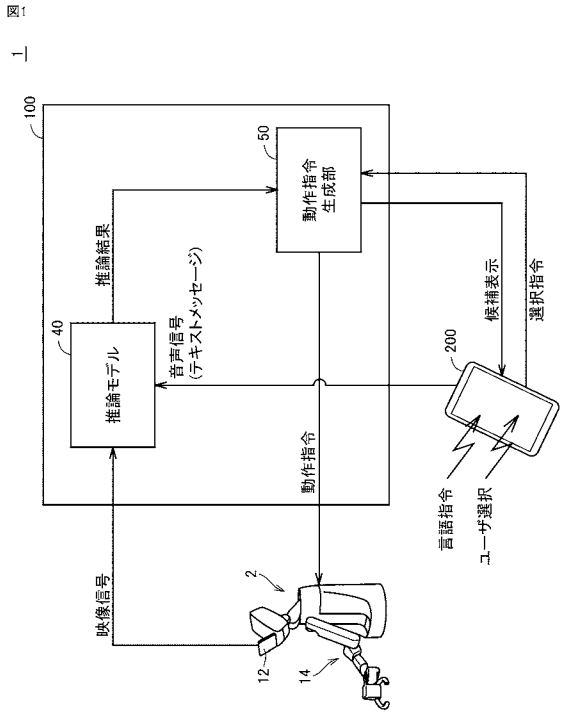


図1
1
100
50
動作指令
生成部
推論結果
音声信号
(テキストメッセージ)
40
推論モデル
200
候補表示
選択指令
動作指令
音語指令
ユーザ選択
映像信号
2
12
14

【図２】

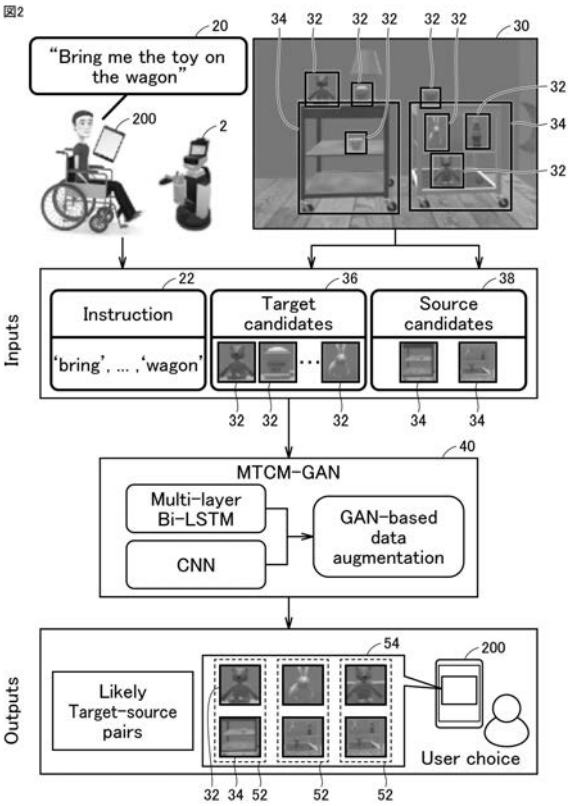


図2
20
"Bring me the toy on the wagon"
200
2
30
32
34
Inputs
22
Instruction
'bring', ... ,'wagon'
36
Target candidates
38
Source candidates
40
MTCM-GAN
Multi-layer Bi-LSTM
CNN
GAN-based data augmentation
Outputs
54
Likely Target-source pairs
User choice
52

【図３】

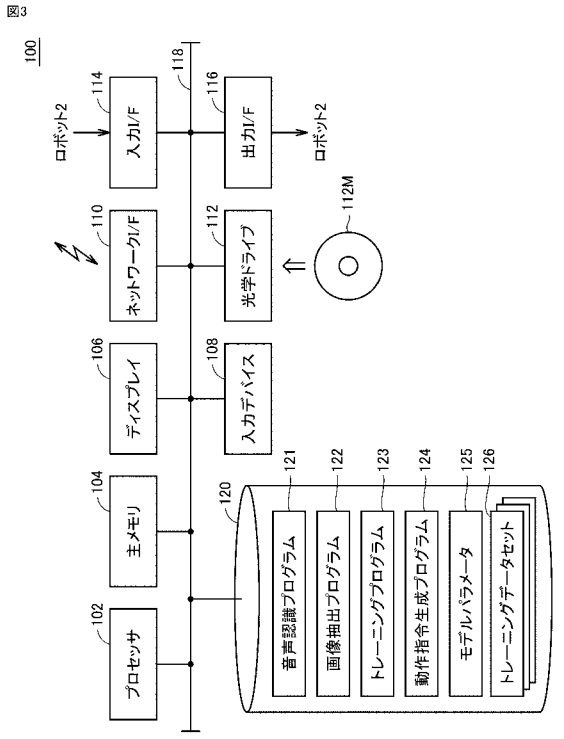


図3
100
ロボット2
入力I/F
114
出力I/F
116
118
ネットワークI/F
110
光学ドライブ
112
112M
ディスプレイ
106
入力デバイス
108
主メモリ
104
プロセッサ
102
120
音声認識プログラム
121
画像抽出プログラム
122
トレーニングプログラム
123
動作指令生成プログラム
124
モデルパラメータ
125
トレーニングデータセット
126

【図４】

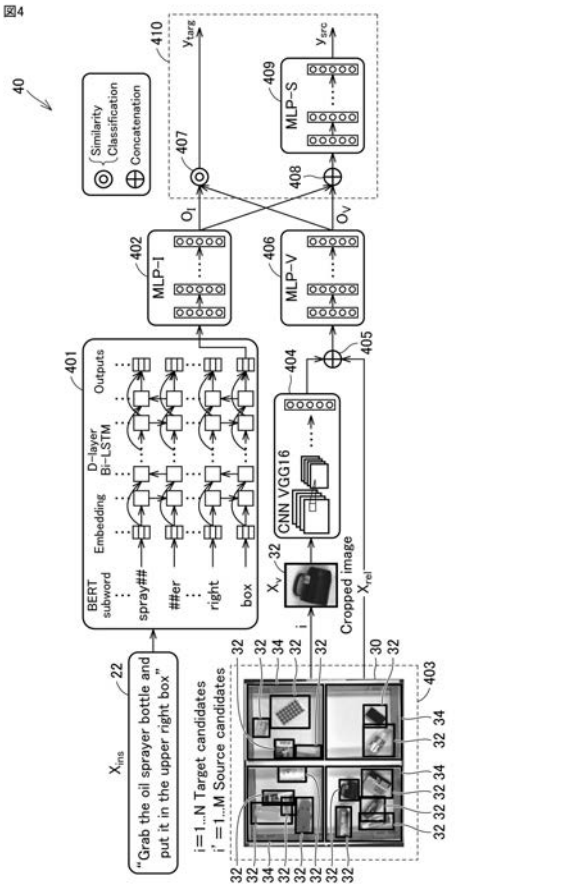
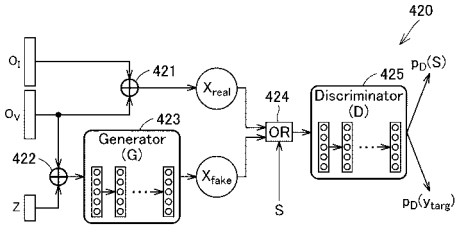


図4
40
Similarity
Classification
Concatenation
410
407
408
409
MLP-S
402
MLP-I
406
MLP-V
401
Outputs
D-layer Bi-LSTM
Embedding
BERT subword
spray##
##er
right
box
404
405
CNN VGG16
Cropped image
22
"Grab the oil sprayer bottle and put it in the upper right box"
i=1...N Target candidates
i'=1...M Source candidates
403

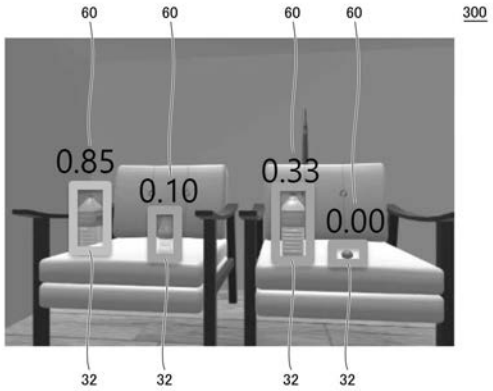
【図５】

図5



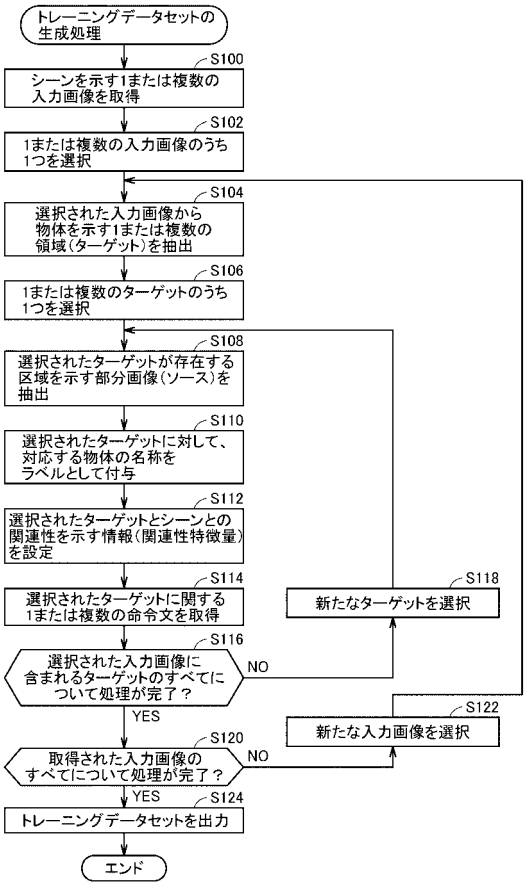
【図６】

図6



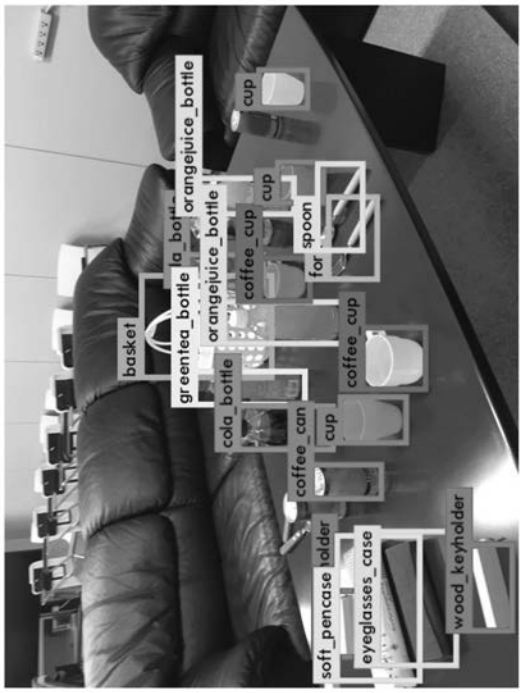
【図７】

図7



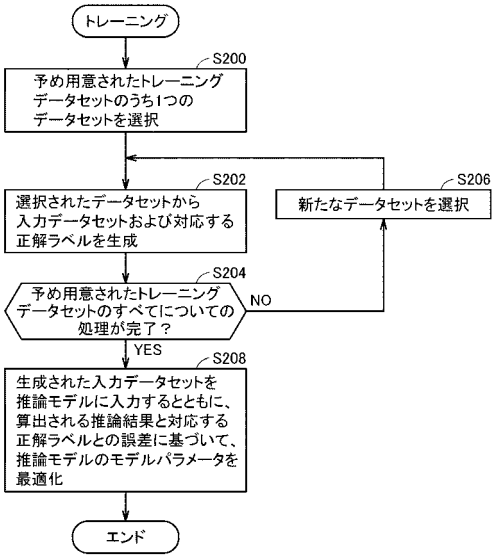
【図８】

図8



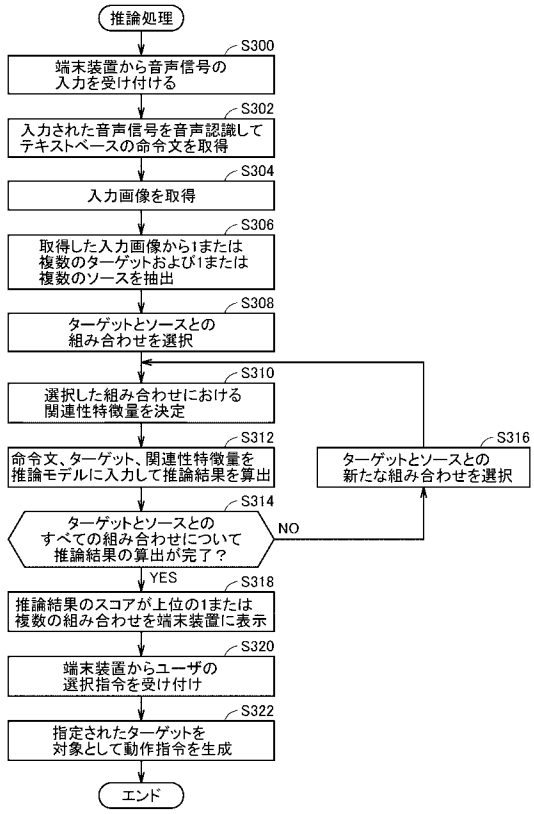
【図９】

図9



【図１０】

図10


推論処理
S300
端末装置から音声信号の
入力を受け付ける
S302
入力された音声信号を音声認識して
テキストベースの命令文を取得
S304
入力画像を取得
S306
取得した入力画像から1または
複数のターゲットおよび1または
複数のソースを抽出
S308
ターゲットとソースとの
組み合わせを選択
S310
選択した組み合わせにおける
関連性特徴量を決定
S312
命令文、ターゲット、関連性特徴量を
推論モデルに入力して推論結果を算出
S314
ターゲットとソースとの
すべての組み合わせについて
推論結果の算出が完了？
NO
YES
S316
ターゲットとソースとの
新たな組み合わせを選択
S318
推論結果のスコアが上位の1または
複数の組み合わせを端末装置に表示
S320
端末装置からユーザの
選択指令を受け付け
S322
指定されたターゲットを
対象として動作指令を生成
エンド

フロントページの続き

Ｆターム(参考) 5E555 AA23 AA46 BA01 BA83 BB01 BC04 BC17 BE17 CA42 CA47
CB45 CB64 CC03 DA31 DB32 DB39 DB53 DC10 DC61 DD11
EA07 EA22 EA25 EA27 FA00
5L096 BA05 HA11 JA11 KA04