(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7538826号
(P7538826)

(45)発行日　令和6年8月22日(2024.8.22)

(24)登録日　令和6年8月14日(2024.8.14)

(51)国際特許分類		FI		
G06T 7/00 (2017.01)		G06T	7/00	350C
G06V 10/82 (2022.01)		G06V	10/82	
G06N 3/04 (2023.01)		G06N	3/04	
G06N 3/08 (2023.01)		G06N	3/08	
G06N 20/00 (2019.01)		G06N	20/00	

請求項の数　14　（全15頁）

(21)出願番号　特願2022-20797(P2022-20797)
(22)出願日　令和4年2月14日(2022.2.14)
(65)公開番号　特開2023-117959(P2023-117959A)
(43)公開日　令和5年8月24日(2023.8.24)
審査請求日　令和5年10月2日(2023.10.2)

(73)特許権者　000001007
キヤノン株式会社
東京都大田区下丸子３丁目３０番２号
(74)代理人　110003281
弁理士法人大塚国際特許事務所
(72)発明者　平戸 康晴
東京都大田区下丸子３丁目３０番２号
キヤノン株式会社内
審査官　秦野 孝一郎

最終頁に続く

(54)【発明の名称】　情報処理装置、情報処理方法

(57)【特許請求の範囲】

【請求項１】

第１演算部が入力データに基づいて抽出した第１特徴ベクトルと、第２演算部が該第１演算部よりも少ない演算量で該入力データに基づいて抽出した第２特徴ベクトルと、の第１差分ベクトルを求める取得手段と、

前記第１差分ベクトルにおける要素の値を変更した第１変更差分ベクトルを生成し、該第１変更差分ベクトルに基づいて前記第２演算部のパラメータを更新する更新手段と

を備え、

前記第１特徴ベクトルにおいて閾値を超える第１要素に対応する前記第１差分ベクトルの要素の値から、前記第１要素に対応する前記第１変更差分ベクトルの要素の値への変化量は、前記第１特徴ベクトルにおいて閾値を超えない第２要素に対応する前記第１差分ベクトルの要素の値から、前記第２要素に対応する前記第１変更差分ベクトルの要素の値への値の変化量よりも大きい

ことを特徴とする情報処理装置。

【請求項２】

前記更新手段は、

前記第１特徴ベクトルにおいて閾値を超える第１要素に対応する前記第１差分ベクトルの要素の値を増加させた前記第１変更差分ベクトルを生成することを特徴とする請求項１に記載の情報処理装置。

【請求項３】

前記更新手段は、
前記第１特徴ベクトルにおいて閾値を超えない第２要素に対応する前記第１差分ベクトルの要素の値を減少させた前記第１変更差分ベクトルを生成することを特徴とする請求項１または２に記載の情報処理装置。
【請求項４】
前記更新手段は、
前記第１変更差分ベクトルにおける要素の値に基づく値がより小さくなるような前記第２演算部のパラメータを求め、前記第２演算部のパラメータを該求めたパラメータに更新することを特徴とする請求項１乃至３の何れか１項に記載の情報処理装置。
【請求項５】
さらに、
前記第１要素に対応する前記第１差分ベクトルの要素の値を増加させる量を増加させる増加手段を備えることを特徴とする請求項２に記載の情報処理装置。
【請求項６】
前記増加手段は、前記第１要素に対応する前記第１差分ベクトルの要素の数が減少しなくなったタイミングで、前記第１要素に対応する前記第１差分ベクトルの要素の値を増加させる量を増加させることを特徴とする請求項５に記載の情報処理装置。
【請求項７】
前記第１演算部は、入力データを階層型ニューラルネットワークに入力し、該階層型ニューラルネットワークの中間層にて該入力データから抽出した特徴ベクトルを前記第１特徴ベクトルとして取得することを特徴とする請求項１乃至６の何れか１項に記載の情報処理装置。
【請求項８】
前記第１演算部が用いる階層型ニューラルネットワークの活性化関数はＲｅＬＵ（Ｒｅｃｔｉｆｉｅｄ Ｌｉｎｅａｒ Ｕｎｉｔ）であることを特徴とする請求項７に記載の情報処理装置。
【請求項９】
前記第２演算部は、前記第１演算部が用いる階層型ニューラルネットワークよりもパラメータ数が少ない階層型ニューラルネットワークに入力データを入力し、該階層型ニューラルネットワークの中間層にて該入力データから抽出した特徴ベクトルを前記第２特徴ベクトルとして取得することを特徴とする請求項１乃至８の何れか１項に記載の情報処理装置。
【請求項１０】
前記第２演算部が用いる階層型ニューラルネットワークの活性化関数はＲｅＬＵ（Ｒｅｃｔｉｆｉｅｄ Ｌｉｎｅａｒ Ｕｎｉｔ）であることを特徴とする請求項９に記載の情報処理装置。
【請求項１１】
前記取得手段はさらに、前記第１演算部が入力データに基づいて抽出した第１中間特徴ベクトルと、前記第２演算部が該第１演算部よりも少ない演算量で該入力データに基づいて抽出した第２中間特徴ベクトルと、の第２差分ベクトルを求め、
前記更新手段は、前記第２差分ベクトルにおける要素の値を変更した第２変更差分ベクトルを生成し、該第２変更差分ベクトルと、前記第１変更差分ベクトルと、に基づいて前記第２演算部のパラメータを更新し、
前記第１中間特徴ベクトルにおいて閾値を超える第３要素に対応する前記第２差分ベクトルの要素の値から、前記第３要素に対応する前記第２変更差分ベクトルの要素の値への変化量は、前記第１中間特徴ベクトルにおいて閾値を超えない第４要素に対応する前記第２差分ベクトルの要素の値から、前記第４要素に対応する前記第２変更差分ベクトルの要素の値への値の変化量よりも大きい
ことを特徴とする請求項１乃至６の何れか１項に記載の情報処理装置。
【請求項１２】

前記更新手段は、
前記第１変更差分ベクトルにおける要素の値と、前記第２変更差分ベクトルにおける要素の値と、に基づく値がより小さくなるような前記第２演算部のパラメータを求め、前記第２演算部のパラメータを該求めたパラメータに更新することを特徴とする請求項１１に記載の情報処理装置。
【請求項１３】
情報処理装置が行う情報処理方法であって、
前記情報処理装置の取得手段が、第１演算部が入力データに基づいて抽出した第１特徴ベクトルと、第２演算部が該第１演算部よりも少ない演算量で該入力データに基づいて抽出した第２特徴ベクトルと、の第１差分ベクトルを求める取得工程と、
前記情報処理装置の更新手段が、前記第１差分ベクトルにおける要素の値を変更した第１変更差分ベクトルを生成し、該第１変更差分ベクトルに基づいて前記第２演算部のパラメータを更新する更新工程と
を備え、
前記第１特徴ベクトルにおいて閾値を超える第１要素に対応する前記第１差分ベクトルの要素の値から、前記第１要素に対応する前記第１変更差分ベクトルの要素の値への変化量は、前記第１特徴ベクトルにおいて閾値を超えない第２要素に対応する前記第１差分ベクトルの要素の値から、前記第２要素に対応する前記第１変更差分ベクトルの要素の値への値の変化量よりも大きい
ことを特徴とする情報処理方法。
【請求項１４】
コンピュータを、請求項１乃至１２の何れか１項に記載の情報処理装置の各手段として機能させるためのコンピュータプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、学習技術に関するものである。
【背景技術】
【０００２】
近年、撮影された画像内に写っているオブジェクトの画像を高度に処理して有用な情報を抽出する特徴抽出技術が多く提案されている。特にその中でも、ディープネット（或いはディープニューラルネット、ディープラーニング、深層学習とも称される）と呼ばれる多階層のニューラルネットワークを用いて、画像中の物体の特徴ベクトルを抽出する特徴抽出技術に関し、盛んに研究が行われている。
【０００３】
ディープネットを用いた特徴抽出技術の隆盛に関しては周知の通りであるが、さらに近年では、特許文献１のような蒸留と呼ばれるディープネット学習手法が注目されている。蒸留とは、学習済のディープネットモデル（教師モデルと呼ばれる）を用いて、別のネットワークアーキテクチャを持つディープネット（生徒モデルと呼ばれる）の学習を行う手法である。一般的に蒸留による学習は、教師モデルの軽量化を目的として行われることが多く、そのため生徒モデルには教師モデルより軽量のネットワークアーキテクチャを用意することが多い。蒸留では、教師モデルが出力する特徴ベクトルを正解ラベルの代わりに用いて生徒モデルを学習させる。従って蒸留による学習では、通常の学習に必要な大量のラベル付きの学習画像というものが必要ない。このような蒸留技術により、教師モデルの知識を生徒モデルに伝搬させることができることが知られている。
【０００４】
蒸留技術で学習された生徒モデルは、教師モデルと略同一の特徴ベクトルを出力することが可能となる。したがって、生徒モデルと教師モデルとでは、ネットワークアーキテクチャやそれに付随するパラメータ等は違うが、両モデルに同じ画像を入力した場合、両モデルからは略同一の特徴ベクトルが出力される。

【０００５】
　また、非特許文献１には、教師モデルが特徴ベクトルを算出する過程で生成する中間特徴ベクトルと生徒モデルが特徴ベクトルを算出する過程で生成する中間特徴ベクトルとが略同一になるように学習することで、より蒸留による学習のしやすさを向上させる技術なども開示されており、研究開発が活発に行われている
【先行技術文献】
【特許文献】
【０００６】
【文献】US10289962
【非特許文献】
【０００７】
【文献】FITNETS: HINTS FOR THIN DEEP NETS Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta & Yoshua Bengio
【発明の概要】
【発明が解決しようとする課題】
【０００８】
　しかしながら、例えばニューラルネットワークのパラメータ数（階層数やニューロン数など）を教師モデルから大幅に減少させた生徒モデルでの蒸留など、蒸留の難易度が上がると、従来の手法では生徒モデルの特徴ベクトルと教師モデルの特徴ベクトルとが略同一にならない場合がある。
【０００９】
　本発明では、難易度の高い蒸留でも、生徒モデルとして動作する演算部から出力される特徴ベクトルと、教師モデルとして動作する演算部から出力される特徴ベクトルと、を略同一に近づけるための学習技術を提供する。
【課題を解決するための手段】
【００１０】
　本発明の一様態は、第１演算部が入力データに基づいて抽出した第１特徴ベクトルと、第２演算部が該第１演算部よりも少ない演算量で該入力データに基づいて抽出した第２特徴ベクトルと、の第１差分ベクトルを求める取得手段と、前記第１差分ベクトルにおける要素の値を変更した第１変更差分ベクトルを生成し、該第１変更差分ベクトルに基づいて前記第２演算部のパラメータを更新する更新手段とを備え、前記第１特徴ベクトルにおいて閾値を超える第１要素に対応する前記第１差分ベクトルの要素の値から、前記第１要素に対応する前記第１変更差分ベクトルの要素の値への変化量は、前記第１特徴ベクトルにおいて閾値を超えない第２要素に対応する前記第１差分ベクトルの要素の値から、前記第２要素に対応する前記第１変更差分ベクトルの要素の値への値の変化量よりも大きいことを特徴とする。
【発明の効果】
【００１１】
　本発明によれば、難易度の高い蒸留でも、生徒モデルとして動作する演算部から出力される特徴ベクトルと、教師モデルとして動作する演算部から出力される特徴ベクトルと、を略同一に近づけることができる。
【図面の簡単な説明】
【００１２】
【図１】情報処理装置１００のハードウェア構成例を示すブロック図。
【図２】情報処理装置１００の機能構成例を示すブロック図。
【図３】情報処理装置１００の動作を示すフローチャート。
【図４】抽出部２０２および抽出部２０３の機能構成例を示すブロック図。
【図５】中間特徴ベクトルと差分ベクトルの一例を示す図。
【図６】情報処理装置１００の機能構成例を示すブロック図。
【発明を実施するための形態】

【００１３】
以下、添付図面を参照して実施形態を詳しく説明する。尚、以下の実施形態は特許請求の範囲に係る発明を限定するものではない。実施形態には複数の特徴が記載されているが、これらの複数の特徴の全てが発明に必須のものとは限らず、また、複数の特徴は任意に組み合わせられてもよい。さらに、添付図面においては、同一若しくは同様の構成に同一の参照番号を付し、重複した説明は省略する。
【００１４】
［第１の実施形態］
本実施形態では、第１演算部が入力データに基づいて抽出した第１特徴ベクトルと、第２演算部が該第１演算部よりも少ない演算量で該入力データに基づいて抽出した第２特徴ベクトルと、の第１差分ベクトルを求め、該第１差分ベクトルにおける要素の値を変更した第１変更差分ベクトルを生成し、該第１変更差分ベクトルに基づいて該第２演算部のパラメータを更新する情報処理装置の一例について説明する。ここで、第１特徴ベクトルにおいて閾値を超える第１要素に対応する第１差分ベクトルの要素の値から、第１要素に対応する第１変更差分ベクトルの要素の値への変化量は、第１特徴ベクトルにおいて閾値を超えない第２要素に対応する第１差分ベクトルの要素の値から、第２要素に対応する第１変更差分ベクトルの要素の値への値の変化量よりも大きい。
【００１５】
先ず、本実施形態に係る情報処理装置１００のハードウェア構成例について、図１のブロック図を用いて説明する。ＣＰＵ１０１は、ＲＡＭ１０３やＲＯＭ１０２に格納されているコンピュータプログラムやデータを用いて各種の処理を実行する。これによりＣＰＵ１０１は、情報処理装置１００全体の動作制御を行うと共に、情報処理装置１００が行うものとして説明する各種の処理を実行もしくは制御する。
【００１６】
ＲＯＭ１０２には、情報処理装置１００の設定データ、情報処理装置１００の起動に係るコンピュータプログラムやデータ、情報処理装置１００の基本動作に係るコンピュータプログラムやデータ、などが格納されている。
【００１７】
ＲＡＭ１０３は、ＲＯＭ１０２や外部記憶装置１０４からロードされたコンピュータプログラムやデータを格納するためのエリア、ＣＰＵ１０１が各種の処理を実行する際に用いるワークエリア、を有する。このようにＲＡＭ１０３は、各種のエリアを適宜提供することができる。
【００１８】
外部記憶装置１０４は、ハードディスクドライブ装置などの大容量情報記憶装置である。外部記憶装置１０４には、ＯＳ（オペレーティングシステム）、情報処理装置１００が行うものとして説明する各種の処理をＣＰＵ１０１に実行もしくは制御させるためのコンピュータプログラムやデータ、などが保存されている。外部記憶装置１０４に保存されているコンピュータプログラムやデータは、ＣＰＵ１０１による制御に従って適宜ＲＡＭ１０３にロードされ、ＣＰＵ１０１による処理対象となる。
【００１９】
なお、外部記憶装置１０４は、情報処理装置１００から着脱可能なフレキシブルディスク（ＦＤ）やＣｏｍｐａｃｔ　Ｄｉｓｃ（ＣＤ）等の光ディスク、磁気や光カード、ＩＣカード、メモリカードなどを含んでもよい。
【００２０】
ＣＰＵ１０１、ＲＯＭ１０２、ＲＡＭ１０３、外部記憶装置１０４は何れもシステムバス１０８に接続されている。また、システムバス１０８にはさらに、入力Ｉ／Ｆ１０５および出力Ｉ／Ｆ１０６が接続されている。
【００２１】
入力Ｉ／Ｆ１０５には入力部１０９が接続されている。入力部１０９は、キーボード、マウス、タッチパネル画面などのユーザインターフェースであり、ユーザが操作すること

で各種の指示をＣＰＵ１０１に対して入力することができる。
【００２２】
出力Ｉ／Ｆ１０６にはモニタ１１０が接続されている。モニタ１１０は、液晶画面やタッチパネル画面を有し、ＣＰＵ１０１による処理結果を画像や文字などでもって表示することができる。なお、モニタ１１０は、画像や文字を投影するプロジェクタなどの投影装置であっても良い。
【００２３】
このような情報処理装置１００には、ＰＣ（パーソナルコンピュータ）、ＷＳ（ワークステーション）、スマートフォン、タブレット端末装置、などのコンピュータ装置が適用可能である。なお、情報処理装置１００に適用可能なハードウェア構成は図１に示した構成に限らず、適宜変形／変更が可能である。
【００２４】
次に、情報処理装置１００の機能構成例について、図２のブロック図を用いて説明する。以下では、図２の機能部を処理の主体として説明するが、実際には該機能部の機能をＣＰＵ１０１に実行もしくは制御させるためのコンピュータプログラムを該ＣＰＵ１０１が実行することで該機能部の機能が実現される。このようなコンピュータプログラムは外部記憶装置１０４に保存されており、ＣＰＵ１０１による制御に従って適宜ＲＡＭ１０３にロードされ、ＣＰＵ１０１により実行される。なお、図２の機能部の１以上をハードウェアで実装しても良い。
【００２５】
データ取得部２０１は外部記憶装置１０４に保存されている画像３０１をＲＡＭ１０３にロード（取得）する。なお、画像３０１は入力データの一例であり、テキストデータ、音声データなどを入力データとしても良い。
【００２６】
抽出部２０２は、画像３０１を、教師モデルであるＣＮＮ（Ｃｏｎｖｏｌｕｔｉｏｎａｌ　Ｎｅｕｔｒａｌ　Ｎｅｔｗｏｒｋ）等のディープネットに入力し、該ディープネットの中間層にて該画像３０１から抽出した第一の特徴ベクトル３０２を取得する。本実施形態では、第一の特徴ベクトル３０２は５１２次元のベクトルであるものとする。５１２次元の第一の特徴ベクトル３０２は、例えば、画像分類システムや顔認証システムでの利用を想定している。
【００２７】
抽出部２０２が用いるディープネットのパラメータ（重み係数など）は既に学習にて取得されたパラメータであり、このパラメータは、以下に説明する本実施形態の学習の過程で変更はされない。
【００２８】
なお、ディープネットは階層型ニューラルネットワークの一例であり、本実施形態に係る抽出部２０２は、他の種類の階層型ニューラルネットワークを用いて画像３０１から第一の特徴ベクトル３０２を取得するようにしても良い。
【００２９】
抽出部２０３は、抽出部２０２が用いたディープネット（ディープネットＡと称する）よりも演算量が少ないディープネット（ディープネットＢと称する）に画像３０１を入力し、該ディープネットＢにおける中間層にて該画像３０１から抽出した第二の特徴ベクトル３０３（５１２次元）を取得する。
【００３０】
ディープネットＢは生徒モデルであり、例えば、ディープネットＡよりもパラメータ数が少ないディープネット（例えば、ディープネットＡよりも中間層の数が少ないディープネット、ディープネットＡよりもニューロン数が少ないディープネット）、である。
【００３１】
本実施形態では、抽出部２０２が用いるディープネットや抽出部２０３が用いるディープネットにおける活性化関数はＲｅＬＵ（Ｒｅｃｔｉｆｉｅｄ　Ｌｉｎｅａｒ　Ｕｎｉｔ）

であるものとするが、以降の説明はこれに限らない。また、抽出部２０２が用いるディープネットや抽出部２０３が用いるディープネットの出力層からは、画像３０１に対応する１０２４次元のベクトルが出力されるものとする。

【００３２】

差分取得部２０４は、抽出部２０２が取得した第一の特徴ベクトル３０２と、抽出部２０３が取得した第二の特徴ベクトル３０３と、の差分ベクトルに基づいて差分値３０６を求める。

【００３３】

取得部２０４ａは、抽出部２０２が取得した第一の特徴ベクトル３０２と、抽出部２０３が取得した第二の特徴ベクトル３０３と、の差分ベクトル３０４を生成する。例えば取得部２０４ａは、第一の特徴ベクトル３０２におけるｉ（１≦ｉ≦５１２）番目の要素をＡｉ、第二の特徴ベクトル３０３におけるｉ番目の要素をＢｉとすると、差分ベクトル３０４のｉ番目の要素Ｃｉの値として$(Ａｉ－Ｂｉ)^{2}$を求める。本実施形態では、要素Ｚの値もＺと称する場合がある。なお、差分ベクトル３０４のｉ番目の要素ＣｉとしてＡｉとＢｉとの非負の差分を求めるのであれば、他の方法を用いて差分ベクトル３０４のｉ番目の要素Ｃｉを求めても良い。

【００３４】

関数適用部２０４ｂは、第一の特徴ベクトル３０２が有する要素のうち閾値ＴＨ１を超える値を有する要素Ａｊ（１≦ｊ≦５１２）を特定し、該要素Ａｊに対応する差分ベクトル３０４の要素Ｃｊの値を増加させた差分ベクトル３０５を生成する。以下では、差分ベクトル３０４における要素Ｃｊの集合を「関数適用領域」と称する。また本実施形態では閾値ＴＨ１を０とする。

【００３５】

例えば関数適用部２０４ｂは、差分ベクトル３０５におけるｊ番目の要素をＤｊとすると、差分ベクトル３０４の要素Ｃｊに以下の（式１）に示す如く関数ｆを適用することでＤｊを求める。

【００３６】

Ｄｊ＝ｆ（Ｃｊ）＝αｘＣｊ　…　（式１）

ここで、αは１以上の実数値を有する重み値であり、関数ｆは要素Ｃｊの値を重み値αに応じて増加させたＤｊを求めるための関数である。なお、要素Ｃｊの値を増加させたＤｊを求めるための関数であれば、関数適用部２０４ｂが要素Ｃｊに適用する関数は（式１）に示した関数ｆに限らず、他の線形関数を用いても良いし、非線形関数を用いても良い。また、同様の目的を達成できるのであれば、関数を使用することに限らない。

【００３７】

つまり、差分ベクトル３０５の生成方法については、「第一の特徴ベクトル３０２において閾値ＴＨ１を超える第１要素に対応する差分ベクトル３０４の要素の値から、該第１要素に対応する差分ベクトル３０５の要素の値への変化量は、第一の特徴ベクトル３０２において閾値ＴＨ１を超えない第２要素に対応する差分ベクトル３０４の要素の値から、該第２要素に対応する差分ベクトル３０５の要素の値への値の変化量よりも大きい」という条件が満たされるのであれば、どのような生成方法を適用しても良い。

【００３８】

例えば、関数適用部２０４ｂは、第一の特徴ベクトル３０２が有する要素のうち閾値ＴＨ１を超えない値を有する要素Ａ’ｋ（１≦ｋ≦５１２）を特定し、該要素Ａ’ｋに対応する差分ベクトル３０４の要素Ｃ’ｋの値を減少させた差分ベクトル３０５を生成するようにしても良い。例えば関数適用部２０４ｂは、差分ベクトル３０４の要素Ｃ’ｋに以下の（式１－１）に示す如く関数ｆ’を適用することで、差分ベクトル３０５におけるｋ番目の要素Ｄｋを求める。

【００３９】

Ｄｋ＝ｆ’（Ｃ’ｋ）＝βｘＣ’ｋ　…　（式１－１）

ここで、βは０＜β＜１を満たす実数値を有する重み値である。この場合、関数適用部

２０４ｂはさらに（式１）を適用しても良いし、適用しなくても良い。以下の表は、第一の特徴ベクトル３０２、第二の特徴ベクトル３０３、差分ベクトル３０４、差分ベクトル３０５の事例を示している。

【００４０】

【表１】

	１次元目	２次元目	３次元目	・・・	510 次元目	511 次元目	512 次元目
第一の特徴ベクトル３０２	0	255	123	・・・	50	0	0
第二の特徴ベクトル３０３	10	30	0	・・・	0	25	230
差分ベクトル３０４	100	50,625	15,129	・・・	2,500	625	52,900
差分ベクトル３０５（α＝６４）	100	3,240,000	968,256	・・・	160,000	625	52,900

【００４１】

ここでは重み値α＝６４としている。例えば表中に明記されている第一の特徴ベクトル３０２の要素（１番目の要素（１次元目の要素）～５１２番目の要素（５１２次元目の要素））のうち要素の値が閾値ＴＨ１＝０を越える要素は２次元目の要素、３次元目の要素、５１０次元目の要素である。よって、関数適用領域に属する要素は差分ベクトル３０４の要素のうち２次元目の要素、３次元目の要素、５１０次元目の要素となり、これらの要素の値に重み値α＝６４を掛けたベクトルが差分ベクトル３０５となる。

【００４２】

算出部２０４ｃは、差分ベクトル３０５の全ての要素の値の合計値を、上記のディープネットの出力層から出力される特徴ベクトルの次元数である１０２４で割った値を差分値３０６として求める。なお、差分ベクトル３０５における要素の値から差分値３０６を求めるための方法は特定の方法に限らない。例えば算出部２０４ｃは、差分ベクトル３０５の全ての要素の値の合計値を差分値３０６として求めても良いし、差分ベクトル３０５において閾値以上の要素の値の合計値や該合計値を１０２４で割った値を差分値３０６として求めても良い。

【００４３】

更新部２０５は、差分値３０６がより小さくなるような「抽出部２０３が用いるディープネットのパラメータ３０７」を誤差逆伝搬法により算出し、現在設定されている「抽出部２０３が用いるディープネットのパラメータ」をパラメータ３０７に更新する。これにより抽出部２０３は、次に入力された画像を、パラメータ３０７に従って再構成されたディープネットＢに入力し、該ディープネットＢにおける中間層にて該画像から抽出した第二の特徴ベクトル３０３（５１２次元）を取得する。パラメータ３０７に従って再構成されたディープネットＢは、再構成前のディープネットＢにおける重み係数などが変わったものとなっている。

【００４４】

増加部２０６は、第一の特徴ベクトル３０２が有する要素のうち閾値ＴＨ１を超える値を有する要素Ａｊに対応する差分ベクトル３０４の要素Ｃｊの値を増加させる量（増加量）をより増加させる。上記の例では、重み値αをより大きくしたり、重み値βをより小さくする。一般には、増加部２０６は、関数適用領域における、差分ベクトル３０４から差分ベクトル３０５への要素の値の増加量が増加するように設定を更新する。または増加部２０６は、非関数適用領域における、差分ベクトル３０４から差分ベクトル３０５への要素の値の減少量がより減少するように設定を更新する。

【００４５】

増加部２０６による重み値αの増加は、関数適用領域に属する要素の数が学習回数が増加しても減少しなくなったタイミングで実施する。ディープネットの活性化関数にＲｅＬＵを用いる場合、値が０以下の入力に対応する出力は全て０であるため、誤差逆伝搬によるパラメータ更新時の学習率が小さいと出力は０のままになり、局所解に陥りやすい。一方、学習率を大きくしてしまうと第二の特徴ベクトル３０３の値が０でない要素に寄与するパラメータの変動も同時に大きくなるため、適切な学習ができない。そのため、学習開

始時における重み値αの値（初期値）は１に設定し、比較的容易に再現できる第一の特徴ベクトル３０２の要素を出力する抽出部２０３のパラメータを最初に取得する。

【００４６】

前述のタイミングで重み値αを増加させることで、関数適用領域に対応する差分値を増加させる。重み値αによって差分値が増加されると、学習は関数適用領域に寄与するパラメータの学習を促進させる。この手順で順次重み値αを増加させることで、最終的に適切な抽出部２０３のパラメータを取得することが可能となる。

【００４７】

なお、増加部２０６による増加量が増加するように設定を更新するタイミングは、上記のタイミングに限らず、例えば、差分値３０６の増減変動で決定しても良い（増加部２０６による減少量が減少するように設定を更新するタイミングについても同様）。つまり、今回求めた差分値３０６の前回からの変化量が閾値未満であれば更新するタイミング、としても良い。また、学習の繰返し回数（試行回数）や学習開始からの経過時間に応じて定期的若しくは不定期的に更新する、としても良い。

【００４８】

このような上記の処理（データ取得部２０１、抽出部２０２、抽出部２０３、取得部２０４ａ、関数適用部２０４ｂ、算出部２０４ｃ、更新部２０５、増加部２０６の処理）を繰り返すことで、「抽出部２０３が用いるディープネット」の学習を行う。

【００４９】

判定部２０７は、学習の終了条件が満たされたか否かを判定する。学習の終了条件は特定の条件に限らない。例えば、判定部２０７は、ユーザが入力部１０９を操作して学習の終了指示を入力した場合には、学習の終了条件が満たされたと判定する。また例えば、判定部２０７は、学習の繰返し回数が規定回数を超えた場合や、学習の開始からの経過時間が規定時間を超えた場合には、学習の終了条件が満たされたと判定する。

【００５０】

次に、上記の情報処理装置１００の動作について、図３のフローチャートに従って説明する。各ステップにおける処理の詳細については上記の通りであるから、ここでは簡単に説明する。

【００５１】

ステップＳ４０１では、データ取得部２０１は外部記憶装置１０４に保存されている画像３０１をＲＡＭ１０３にロード（取得）する。ステップＳ４０２では、抽出部２０２は、ステップＳ４０１で取得した画像３０１をディープネットＡに入力し、該ディープネットＡにおける中間層にて該画像３０１から抽出した第一の特徴ベクトル３０２を取得する。

【００５２】

ステップＳ４０３では、抽出部２０３は、ステップＳ４０１で取得した画像３０１をディープネットＢに入力し、該ディープネットＢにおける中間層にて該画像３０１から抽出した第二の特徴ベクトル３０３を取得する。

【００５３】

ステップＳ４０４では、取得部２０４ａは、ステップＳ４０２で取得した第一の特徴ベクトル３０２と、ステップＳ４０３で取得した第二の特徴ベクトル３０３と、の差分ベクトル３０４を生成する。

【００５４】

ステップＳ４０５では、関数適用部２０４ｂは、第一の特徴ベクトル３０２が有する要素のうち閾値ＴＨ１を超える値を有する要素Ａｊを特定し、該要素Ａｊに対応する差分ベクトル３０４の要素Ｃｊの値を増加させた差分ベクトル３０５を生成する。

【００５５】

ステップＳ４０６では、算出部２０４ｃは、ステップＳ４０５で生成した差分ベクトル３０５の全ての要素の値の合計値を１０２４で割った値を差分値３０６として求める。

【００５６】

ステップＳ４０７では、更新部２０５は、ステップＳ４０６で求めた差分値３０６がよ

り小さくなるような「抽出部２０３が用いるディープネットのパラメータ３０７」を誤差逆伝搬法により算出する。そして更新部２０５は、現在設定されている「抽出部２０３が用いるディープネットのパラメータ」をパラメータ３０７に更新する。

【００５７】

ステップＳ４０８では、増加部２０６は、重み値αを更新するタイミングか否かを判断する。この判断の結果、重み値αを更新するタイミングである場合には、処理はステップＳ４０９に進み、重み値αを更新するタイミングではない場合には、処理はステップＳ４１０に進む。

【００５８】

例えば、関数適用領域に属する要素の数が１００回の学習を通して連続で変化しないという条件が満たされた場合、重み値αを更新するタイミングであると判断し、該条件が満たされていない場合には、重み値αを更新するタイミングではないと判断する。

【００５９】

ステップＳ４０９では、増加部２０６は、重み値αをより大きくする。ここでは、重み値αの初期値は１とし、１回で増加させる重み値αの増分を６４とする。ステップＳ４１０では、判定部２０７は、学習の終了条件が満たされたか否かを判定する。この判定の結果、学習の終了条件が満たされた場合には、図３のフローチャートに従った処理は終了し、学習の終了条件が満たされていない場合には、処理はステップＳ４０１に進む。

【００６０】

なお、ＣＰＵ１０１は、図３のフローチャートに従った学習処理によって得られる「抽出部２０３が用いるディープネットのパラメータ３０７」を外部記憶装置１０４に保存するようにしても良いし、ネットワークを介して外部の装置に対して送信しても良い。また、ＣＰＵ１０１は、「抽出部２０３が用いるディープネットのパラメータ３０７」を画像や文字などでもってモニタ１１０に表示させても良い。このように、「抽出部２０３が用いるディープネットのパラメータ３０７」の出力先や出力形態については特定の出力先、特定の出力形態に限らない。

【００６１】

このように、本実施形態によれば、蒸留の際に関数適用領域の差分に重み値を付与し、学習の進捗に応じて重み値を増加させる。これにより、従来の手法では実現が困難であった難易度の高い蒸留であっても、生徒モデルの特徴ベクトルと教師モデルの特徴ベクトルとを略同一にすることが可能となる。

【００６２】

［第２の実施形態］

以下では第１の実施形態との差分について説明し、以下で特に触れない限りは第１の実施形態と同様であるものとする。本実施形態に係る抽出部２０２および抽出部２０３の機能構成例について、図４のブロック図を用いて説明する。

【００６３】

本実施形態に係る抽出部２０２は、第一の前半抽出部５０１と第一の後半抽出部５０２とを有する。第一の前半抽出部５０１は、入力された画像３０１をディープネットＡに入力し、該ディープネットＡにおける中間層Ａにて該画像３０１から抽出した第一の中間特徴ベクトル５０５を取得する。中間層Ａとは、ディープネットＡにおける入力層と「第一の特徴ベクトル３０２を出力する中間層」との間における中間層である。第一の後半抽出部５０２は、該第一の中間特徴ベクトル５０５を入力として中間層Ａ以降の各層の演算を行うことで第一の特徴ベクトル３０２を生成する。

【００６４】

本実施形態に係る抽出部２０３は、第二の前半抽出部５０３と第二の後半抽出部５０４とを有する。第二の前半抽出部５０３は、入力された画像３０１をディープネットＢに入力し、該ディープネットＢにおける中間層Ｂにて該画像３０１から抽出した第二の中間特徴ベクトル５０６を取得する。中間層Ｂとは、ディープネットＢにおける入力層と「第二の特徴ベクトル３０３を出力する中間層」との間における中間層である。第二の後半抽出

部５０４は、該第二の中間特徴ベクトル５０６を入力として中間層Ｂ以降の各層の演算を行うことで第二の特徴ベクトル３０３を生成する。

【００６５】

第一の特徴ベクトル３０２は、学習済みのディープネットから取得されるため、正解ラベルと近しい値になっている。そのため、正解ラベルのベクトルの要素の値がほとんど０で占められている場合は第一の特徴ベクトル３０２も同様に多くの要素の値が０になり、本実施形態のように閾値ＴＨ１を０とする場合、関数適用領域に属する要素がほとんど存在しない場合がある。

【００６６】

一方、第一の中間特徴ベクトル５０５は、正解ラベルと近しい値になるように第一の特徴ベクトル３０２を取得するために画像３０１からさまざまな特徴を取得したベクトルになっており、ベクトルの要素の値の多くが非ゼロになる。そのため、第一の中間特徴ベクトル５０５については関数適用領域に属する要素が多く、中間特徴ベクトルを用いて蒸留を行えば上記の学習をより効果的に実行することができる。

【００６７】

図５を用いて中間特徴ベクトルと差分ベクトルについて説明する。図５は、２次元データが９つ存在する３次元の中間特徴ベクトルの事例を示している。黒く塗りつぶされた領域は値が０であることを示し、白く塗りつぶされた領域は値が非ゼロであることを示している。

【００６８】

図５（ａ）は、第一の中間特徴ベクトル５０５の事例を示しており、学習済みのディープネット（抽出部２０２）からの中間特徴ベクトルは画像３０１のさまざまな特徴を取得しているため、多くが非ゼロになっている。

【００６９】

図５（ｂ）は、学習途中のディープネット（抽出部２０３）からの第二の中間特徴ベクトル５０６の事例を示しており、３つの２次元データで全てゼロの値となっており、その他の２次元データが第一の中間特徴ベクトル５０５と同等の値になっている。

【００７０】

図５（ｃ）は、第一の中間特徴ベクトル５０５と第二の中間特徴ベクトル５０６の差分ベクトルの事例を示しており、３つの２次元データで差分が発生している。差分が発生している領域は全て関数適用領域であるため、重み値αの値を増加させることで３つの２次元データに寄与するパラメータの学習を促進することができる。

【００７１】

本実施形態に係る情報処理装置１００の機能構成例について、図６のブロック図を用いて説明する。抽出部２０２は、画像３０１から第一の特徴ベクトル３０２と第一の中間特徴ベクトル５０５を取得する。抽出部２０３は、画像３０１から第二の特徴ベクトル３０３と第二の中間特徴ベクトル５０６を取得する。

【００７２】

差分取得部２０４は、第一の特徴ベクトル３０２と第二の特徴ベクトル３０３から第１の実施形態と同様にして差分値７０１を取得し、第一の中間特徴ベクトル５０５と第二の中間特徴ベクトル５０６から第１の実施形態と同様にして差分値７０２を取得する。２つのベクトルから差分値を求める為の処理については第１の実施形態で説明したように様々な処理がある。よって、差分値７０１を求める為の処理と、差分値７０２を求める為の処理と、は同じ処理であっても良いし、異なる処理であっても良いし、同じ処理であっても閾値や重み値αを変えても良い。

【００７３】

統合部７０３は、差分値７０１と差分値７０２の合計値を差分値３０６として求める。以降は第１の実施形態と同様の処理を行ってパラメータ３０７を求め、抽出部２０３のパラメータを該求めたパラメータ３０７で更新する。

【００７４】

また、差分値７０１を求める為に用いた閾値や重み値αの更新タイミング、差分値７０２を求める為に用いた閾値や重み値αの更新タイミング、は同じであっても良いし、異なっていても良い。また、差分値７０１を求める為に用いた重み値αの増加量、差分値７０２を求める為に用いた重み値αの増加量、は同じであっても良いし、異なっていても良い。

【００７５】

本実施形態では、図３のフローチャートにおいて、ステップＳ４０２にて抽出部２０２は、画像３０１から第一の特徴ベクトル３０２と第一の中間特徴ベクトル５０５を取得する。そしてステップＳ４０３にて、抽出部２０３は、画像３０１から第二の特徴ベクトル３０３と第二の中間特徴ベクトル５０６を取得する。

【００７６】

そして差分取得部２０４は、ステップＳ４０４～ステップＳ４０６の処理で第１の実施形態と同様に第一の特徴ベクトル３０２と第二の特徴ベクトル３０３から差分値７０１を取得する。また差分取得部２０４は、ステップＳ４０４～ステップＳ４０６の処理と同様の処理により、第一の中間特徴ベクトル５０５と第二の中間特徴ベクトル５０６から差分値７０２を取得する。そしてステップＳ４０６では、統合部７０３は、差分値７０１と差分値７０２の合計値を差分値３０６として求める。その他のステップにおける処理は第１の実施形態と同様である。

【００７７】

このように、本実施形態によれば、従来の手法では実現が困難であった難易度の高い蒸留であっても、中間特徴ベクトルを利用する蒸留の場合はより高い精度で、生徒モデルの特徴ベクトルと教師モデルの特徴ベクトルとを略同一にすることが可能となる。

【００７８】

また、上記の各実施形態で使用した数値、処理タイミング、処理順、処理の主体、データ（情報）の送信先／送信元／格納場所などは、具体的な説明を行うために一例として挙げたもので、このような一例に限定することを意図したものではない。

【００７９】

また、以上説明した各実施形態の一部若しくは全部を適宜組み合わせて使用しても構わない。また、以上説明した各実施形態の一部若しくは全部を選択的に使用しても構わない。

【００８０】

（その他の実施形態）

本発明は、上述の実施形態の１以上の機能を実現するプログラムを、ネットワーク又は記憶媒体を介してシステム又は装置に供給し、そのシステム又は装置のコンピュータにおける１つ以上のプロセッサがプログラムを読出し実行する処理でも実現可能である。また、１以上の機能を実現する回路（例えば、ＡＳＩＣ）によっても実現可能である。

【００８１】

発明は上記実施形態に制限されるものではなく、発明の精神及び範囲から離脱することなく、様々な変更及び変形が可能である。従って、発明の範囲を公にするために請求項を添付する。

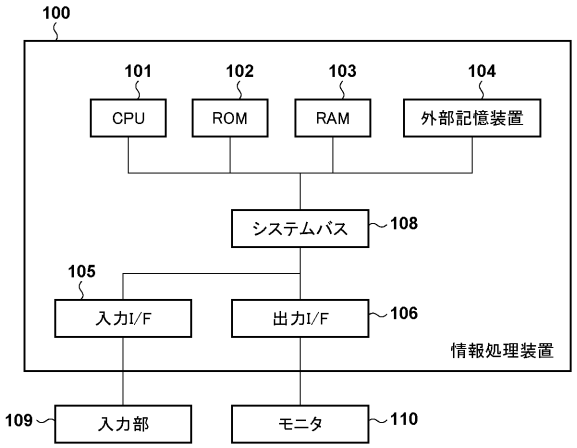
【符号の説明】

【００８２】

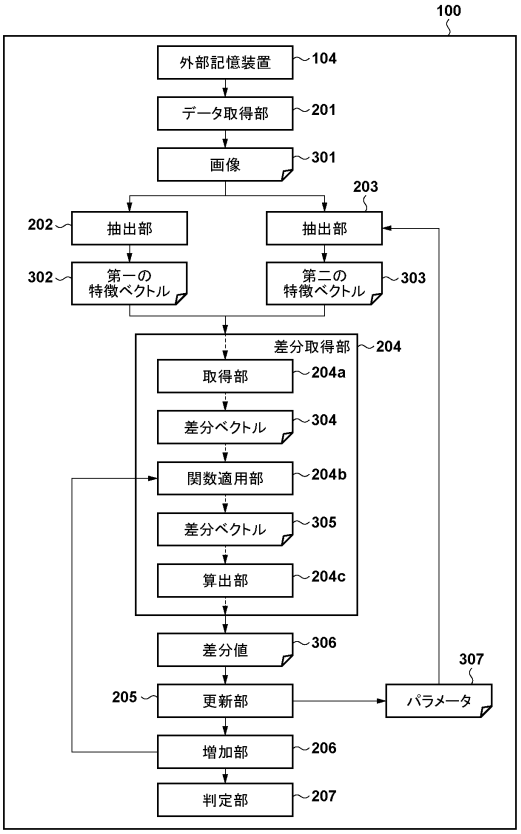
２０１：データ取得部　２０２：抽出部　２０３：抽出部　２０４：差分取得部　２０４ａ：取得部　２０４ｂ：関数適用部　２０４ｃ：算出部　２０５：更新部　２０６：増加部　２０７：判定部

【図面】

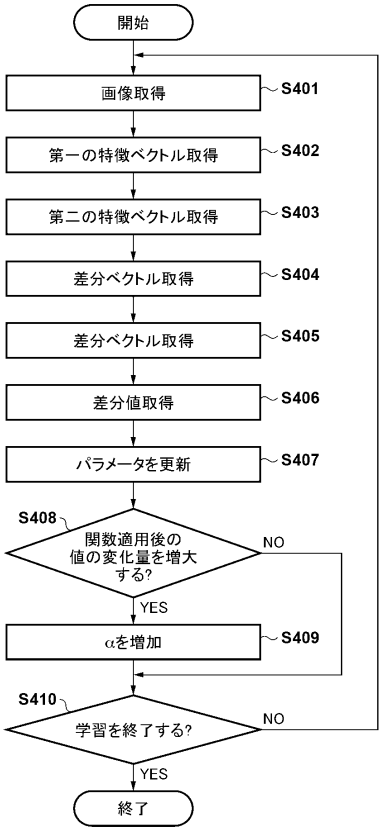
【図１】


100
101
CPU
102
ROM
103
RAM
104
外部記憶装置
システムバス
108
105
入力I/F
出力I/F
106
情報処理装置
109
入力部
モニタ
110

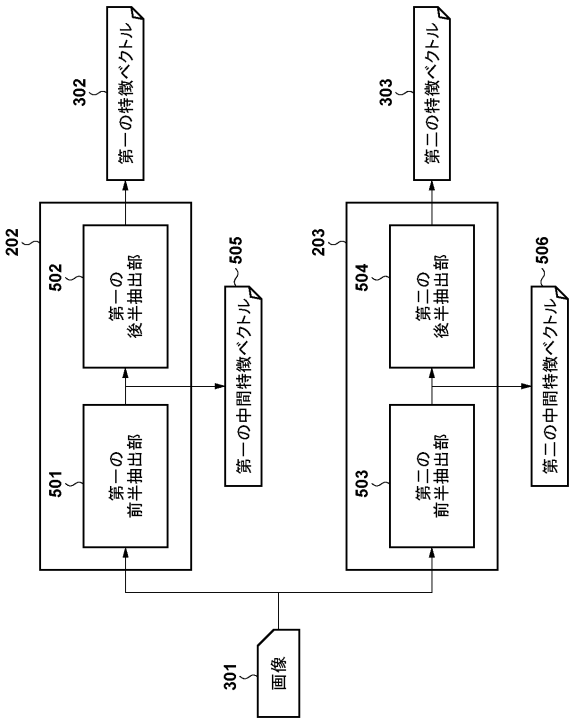
【図２】


100
外部記憶装置
104
データ取得部
201
画像
301
203
202
抽出部
抽出部
302
第一の特徴ベクトル
第二の特徴ベクトル
303
差分取得部
204
取得部
204a
差分ベクトル
304
関数適用部
204b
差分ベクトル
305
算出部
204c
差分値
306
307
205
更新部
パラメータ
増加部
206
判定部
207

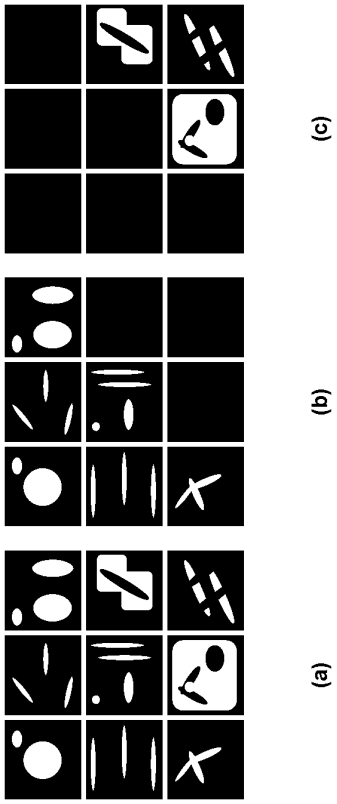
【図３】


開始
画像取得
S401
第一の特徴ベクトル取得
S402
第二の特徴ベクトル取得
S403
差分ベクトル取得
S404
差分ベクトル取得
S405
差分値取得
S406
パラメータを更新
S407
S408
関数適用後の値の変化量を増大する?
NO
YES
αを増加
S409
S410
学習を終了する?
NO
YES
終了

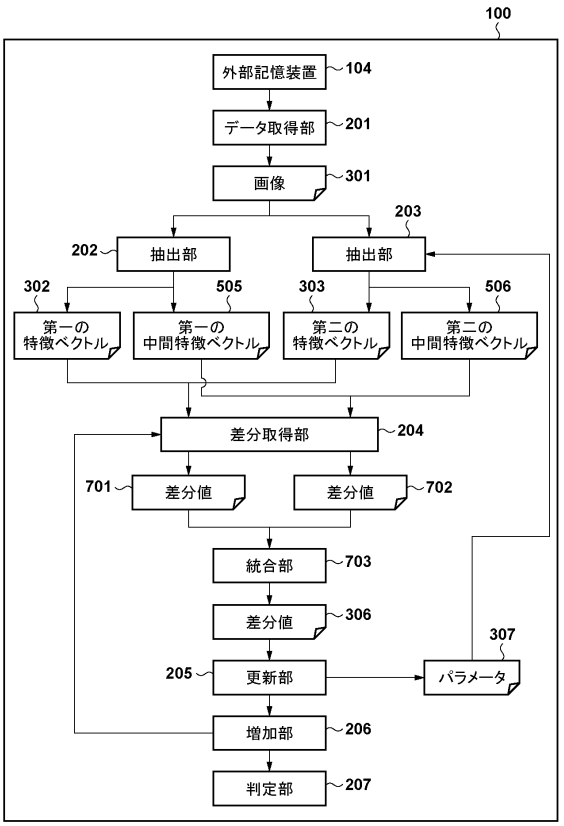
【図４】


302
第一の特徴ベクトル
303
第二の特徴ベクトル
202
502
第一の後半抽出部
505
第一の中間特徴ベクトル
203
504
第二の後半抽出部
506
第二の中間特徴ベクトル
501
第一の前半抽出部
503
第二の前半抽出部
301
画像

【図５】


(c)
(b)
(a)

【図６】


100
外部記憶装置
104
データ取得部
201
画像
301
203
202
抽出部
抽出部
302
505
303
506
第一の特徴ベクトル
第一の中間特徴ベクトル
第二の特徴ベクトル
第二の中間特徴ベクトル
差分取得部
204
701
差分値
差分値
702
統合部
703
差分値
306
307
205
更新部
パラメータ
増加部
206
判定部
207

フロントページの続き

(56)参考文献 特開２０１９－１５２９７７（ＪＰ，Ａ）
特開２０２２－２０２３（ＪＰ，Ａ）
Tao Wang, Li Yuan, Xiaopeng Zhang, Jiashi Feng，Distilling Object Detectors With Fine-Grained Feature Imitation，2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)，米国，IEEE，2019年06月15日，pp.4928-4937，https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8953432
Ivan Lazarevich et al.，Post-training deep neural network pruning via layer-wise calibration，2021 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)，米国，IEEE，2021年10月11日，pp.798-805，https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9607584

(58)調査した分野 (Int.Cl.，ＤＢ名)
Ｇ０６Ｔ　７／００－７／９０
Ｇ０６Ｖ　１０／００－２０／９０
Ｇ０６Ｎ　３／０４
Ｇ０６Ｎ　３／０８
Ｇ０６Ｎ　２０／００