(19)日本国特許庁(JP)

(12)特許公報(B1)

(11)特許番号
特許第7682408号
(P7682408)

(45)発行日　令和7年5月23日(2025.5.23)

(24)登録日　令和7年5月15日(2025.5.15)

(51)国際特許分類
Ｇ０６Ｔ　7/00 (2017.01)

ＦＩ
Ｇ０６Ｔ　7/00　３５０Ｂ

請求項の数　16　（全20頁）

(21)出願番号　特願2024-569411(P2024-569411)
(86)(22)出願日　令和6年4月23日(2024.4.23)
(86)国際出願番号　PCT/JP2024/015829
審査請求日　令和6年11月22日(2024.11.22)
早期審査対象出願

(73)特許権者　000006013
三菱電機株式会社
東京都千代田区丸の内二丁目７番３号
(74)代理人　100118762
弁理士　高村 順
(72)発明者　飯野 晋
東京都千代田区丸の内二丁目７番３号
三菱電機株式会社内
(72)発明者　内藤 美里
東京都千代田区丸の内二丁目７番３号
三菱電機株式会社内
(72)発明者　山田 和彦
東京都千代田区丸の内二丁目７番３号
三菱電機株式会社内
審査官　山田 辰美

最終頁に続く

(54)【発明の名称】　画像認識システム、画像認識装置、文章特徴抽出装置、プログラムおよび画像認識方法

(57)【特許請求の範囲】

【請求項１】

認識対象の画像データから検出対象を検出する画像認識システムであって、

文章特徴抽出装置と、画像認識装置と、を備え、

前記文章特徴抽出装置は、

前記検出対象を指定するための自然言語の語句を取得する文章取得部と、

前記自然言語の語句から画像の特徴量と比較可能な特徴量である文章特徴を抽出する文章特徴抽出部と、

前記文章特徴を送信する文章特徴送信部と、

を有し、

前記画像認識装置は、

前記文章特徴抽出装置から前記文章特徴を受信する文章特徴受信部と、

前記文章特徴受信部が受信した前記文章特徴を記憶する文章特徴記憶部と、

前記画像データを取得する画像取得部と、

前記画像データから画像内で前記検出対象が存在する領域の候補である候補領域、および、前記候補領域内の画像の特徴量である画像特徴を抽出する画像解析部と、

前記文章特徴記憶部に記憶された前記文章特徴と、前記候補領域に対応する前記画像特徴とを比較して、比較結果を出力する特徴比較部と、

前記比較結果に基づいて、前記検出対象の検出結果を生成する検出結果生成部と、

を有し、

前記文章特徴抽出部の処理を、前記画像認識装置とは別の装置である前記文章特徴抽出装置で実行する
ことを特徴とする画像認識システム。
【請求項２】
前記画像解析部は、前記画像データから前記自然言語の語句で指定された前記検出対象を検出するための学習済モデルのうちの一部であって、出力が自然言語の入力に依存しない第１の部分モデルを実行する
ことを特徴とする請求項１に記載の画像認識システム。
【請求項３】
前記文章特徴抽出部は、前記画像データから前記自然言語の語句で指定された前記検出対象を検出するための学習済モデルの一部であって、出力が画像データの入力に依存しない第２の部分モデルを実行する
ことを特徴とする請求項１に記載の画像認識システム。
【請求項４】
前記特徴比較部は、前記画像データから前記自然言語の語句で指定された前記検出対象を検出するための学習済モデルの一部であって、自然言語の入力に依存しない第１の部分モデルの出力と、画像データの入力に依存しない第２の部分モデルの出力とを入力として受け付けて、前記比較結果を出力する第３の部分モデルを実行する
ことを特徴とする請求項１に記載の画像認識システム。
【請求項５】
前記文章特徴抽出装置は、
前記文章取得部が新たな前記自然言語の語句を取得する毎に、前記文章特徴抽出部によって、新たな前記自然言語の語句から前記文章特徴を抽出し、抽出した前記文章特徴を前記文章特徴送信部によって送信する
ことを特徴とする請求項１に記載の画像認識システム。
【請求項６】
認識対象の画像データから検出対象を検出する画像認識システムであって、
画像認識装置と、前記画像認識装置よりも演算処理能力が高い文章特徴抽出装置と、を備え、
前記文章特徴抽出装置は、
前記検出対象を指定するための自然言語の語句を取得する文章取得部と、
前記自然言語の語句から画像の特徴量と比較可能な特徴量である文章特徴を抽出する文章特徴抽出部と、
前記文章特徴を送信する文章特徴送信部と、
を有し、
前記画像認識装置は、
前記文章特徴抽出装置から前記文章特徴を受信する文章特徴受信部と、
前記画像データを取得する画像取得部と、
前記画像データから画像内で前記検出対象が存在する領域の候補である候補領域、および、前記候補領域内の画像の特徴量である画像特徴を抽出する画像解析部と、
前記文章特徴と、前記候補領域に対応する前記画像特徴とを比較して、比較結果を出力する特徴比較部と、
前記比較結果に基づいて、前記検出対象の検出結果を生成する検出結果生成部と、
を有し、
前記文章特徴抽出部の処理を、前記画像認識装置とは別の装置である前記文章特徴抽出装置で実行する
ことを特徴とする画像認識システム。
【請求項７】
認識対象の画像データから検出対象を検出する画像認識装置であって、
画像の特徴量と比較可能な特徴量であって前記検出対象を指定するための自然言語の語

句から抽出された文章特徴を前記画像認識装置の外部から受信する文章特徴受信部と、
前記文章特徴受信部が受信した前記文章特徴を記憶する文章特徴記憶部と、
前記画像データを取得する画像取得部と、
前記画像データから画像内で前記検出対象が存在する領域の候補である候補領域、および、前記候補領域内の画像の特徴量である画像特徴を抽出する画像解析部と、
前記文章特徴記憶部に記憶された前記文章特徴と、前記候補領域に対応する前記画像特徴とを比較して、比較結果を出力する特徴比較部と、
前記比較結果に基づいて、前記検出対象の検出結果を生成する検出結果生成部と、
を備える
ことを特徴とする画像認識装置。

【請求項８】
前記画像解析部は、前記画像データから前記自然言語の語句で指定された前記検出対象を検出するための学習済モデルのうちの一部であって、出力が自然言語の入力に依存しない第１の部分モデルを実行する
ことを特徴とする請求項７に記載の画像認識装置。

【請求項９】
前記特徴比較部は、前記学習済モデルの一部であって、前記第１の部分モデルの出力と、前記学習済モデルの一部であって画像データの入力に依存しない第２の部分モデルの出力とを入力として受け付けて、前記比較結果を出力する第３の部分モデルを実行する
ことを特徴とする請求項８に記載の画像認識装置。

【請求項１０】
画像データから検出する検出対象を指定するための自然言語の語句を取得する文章取得部と、
前記自然言語の語句から画像の特徴量と比較可能な特徴量である文章特徴を抽出する文章特徴抽出部と、
前記文章特徴を、前記画像データから前記検出対象を検出する画像認識装置に送信する文章特徴送信部と、
を備え、
新たな前記自然言語の語句を入力されたとき、前記文章特徴抽出部は、前記画像データから前記自然言語の語句で指定された前記検出対象を検出するための学習済モデルの一部であって、出力が画像データの入力に依存しない第２の部分モデルを実行することによって前記文章特徴を抽出し、前記文章特徴送信部は、抽出された前記文章特徴を前記画像認識装置に送信し、
前記検出対象の変更がなく新たな前記自然言語の語句が入力されないとき、新たな前記画像データが取得されても前記第２の部分モデルの実行を省略する
ことを特徴とする文章特徴抽出装置。

【請求項１１】
コンピュータを、認識対象の画像データから検出対象を検出する画像認識装置として機能させるためのプログラムであって、
画像の特徴量と比較可能な特徴量であって、当該コンピュータとは異なるコンピュータにおいて前記検出対象を指定するための自然言語の語句から抽出された文章特徴を受信するステップと、
受信した前記文章特徴を前記画像認識装置として機能させるコンピュータ内に記憶するステップと、
前記画像データを取得するステップと、
前記画像データから画像内で前記検出対象が存在する領域の候補である候補領域、および、前記候補領域内の画像の特徴量である画像特徴を抽出するステップと、
前記画像認識装置として機能させるコンピュータ内に記憶された前記文章特徴と、前記候補領域に対応する前記画像特徴とを比較して、比較結果を出力するステップと、
前記比較結果に基づいて、前記検出対象の検出結果を生成するステップと、

を前記画像認識装置として機能させるコンピュータに実行させる
ことを特徴とするプログラム。
【請求項１２】
コンピュータを、文章特徴抽出装置として機能させるためのプログラムであって、
画像データから検出する検出対象を指定するための自然言語の語句を取得するステップと、
新たな前記自然言語の語句を入力されたとき、前記画像データから前記自然言語の語句で指定された前記検出対象を検出するための学習済モデルの一部であって、出力が画像データの入力に依存しない第２の部分モデルを実行することによって前記自然言語の語句から画像の特徴量と比較可能な特徴量である文章特徴を抽出するステップと、
前記文章特徴を、前記文章特徴抽出装置として機能させるコンピュータとは異なるコンピュータであって、前記画像データから前記検出対象を検出する画像認識装置として機能するコンピュータに送信するステップと、
を前記文章特徴抽出装置として機能させるコンピュータに実行させ、
前記検出対象の変更がなく新たな前記自然言語の語句が入力されないとき、新たな前記画像データが取得されても前記文章特徴を抽出するステップを省略して前記第２の部分モデルを実行しない
ことを特徴とするプログラム。
【請求項１３】
認識対象の画像データから検出対象を検出する画像認識方法であって、
文章特徴抽出装置が、前記検出対象を指定するための自然言語の語句を取得するステップと、
前記文章特徴抽出装置が、前記自然言語の語句から画像の特徴量と比較可能な特徴量である文章特徴を抽出するステップと、
前記文章特徴抽出装置が、抽出した前記文章特徴を画像認識装置に送信するステップと、
前記画像認識装置が、前記文章特徴抽出装置から前記文章特徴を受信するステップと、
受信した前記文章特徴を前記画像認識装置の有する文章特徴記憶部に記憶するステップと、
前記画像認識装置が、前記画像データを取得するステップと、
前記画像認識装置が、前記画像データから画像内で前記検出対象が存在する領域の候補である候補領域、および、前記候補領域内の画像の特徴量である画像特徴を抽出するステップと、
前記画像認識装置が、前記文章特徴記憶部に記憶された前記文章特徴と、前記候補領域に対応する前記画像特徴とを比較して、比較結果を生成するステップと、
前記画像認識装置が、前記比較結果に基づいて、前記検出対象の検出結果を生成するステップと、
を含む
ことを特徴とする画像認識方法。
【請求項１４】
コンピュータを、文章特徴抽出装置として機能させるためのプログラムであって、
画像データから検出する検出対象を指定するための自然言語の語句を、新たな前記自然言語の語句が入力される毎に取得するステップと、
前記自然言語の語句から画像の特徴量と比較可能な特徴量である文章特徴を抽出するステップと、
前記文章特徴を、前記文章特徴抽出装置として機能させるコンピュータとは異なるコンピュータであって、受信した前記文章特徴を記憶し、記憶された前記文章特徴を用いて前記画像データから前記検出対象を検出する画像認識装置として機能するコンピュータに送信するステップと、
を前記文章特徴抽出装置として機能させるコンピュータに実行させる
ことを特徴とするプログラム。
【請求項１５】

請求項１１に記載のプログラムを実行する画像認識装置を備えた画像認識システム。
【請求項１６】
請求項１２または１４に記載のプログラムを実行する文章特徴抽出装置を備えた画像認識システム。
【発明の詳細な説明】
【技術分野】
【０００１】
本開示は、画像データから自然言語の語句で指定された検出対象を検出する画像認識システム、画像認識装置、文章特徴抽出装置、プログラムおよび画像認識方法に関する。
【背景技術】
【０００２】
画像データから検出対象を検出する画像認識技術では、人工知能の発達に伴って高精度化が著しく、高度なセキュリティシステムなどの実現が期待されている。人工知能で用いられる学習済モデルは、従来の画像認識技術と比較して計算量が非常に多く、例えば監視カメラのように計算能力に制限のある機器では実装することが困難であった。
【０００３】
上述の問題に対して、非特許文献１には、Grounding DINO（self-DIstillation with NO labels：ラベル無自己蒸留）と呼ばれる技術が開示されている。Grounding DINOは、従来の深層ニューラルネットワーク（ＤＮＮ）と比較して１０倍から１００倍の膨大な画像データと、ペアとなる自然言語で示された語句とを入力として、事前にＤＮＮモデルを学習することで、より広い範囲の物体カテゴリの検出に対応することが可能な技術である。また、Grounding DINOでは、推論時には検出対象となる物体カテゴリを自然言語の語句で指定することにより、ＤＮＮモデルの再学習なしに検出対象となる物体カテゴリを変更したり追加したりすることが可能である。
【先行技術文献】
【非特許文献】
【０００４】
【文献】Shilong Liu，他10名，“Grounding DINO：Marrying DINO with Grounded Pre-Training for Open-Set Object Detection”，[online]，[令和５年１２月１日検索]，インターネット＜URL:https://arxiv.org/pdf/2303.05499.pdf＞
【発明の概要】
【発明が解決しようとする課題】
【０００５】
しかしながら、非特許文献１に開示された技術でも、依然として、学習済モデルを用いて、自然言語で入力される語句で指定された検出対象を画像データから検出する推論処理の計算量が非常に大きく、画像認識装置の計算能力に制限がある場合には実装することが困難であるという問題があった。
【０００６】
本開示は、上記に鑑みてなされたものであって、画像認識装置の計算能力に制限がある場合であっても、学習済モデルを用いて、自然言語で入力される語句で指定された検出対象を画像データから検出する推論処理を実装することが可能な画像認識システムを得ることを目的とする。
【課題を解決するための手段】
【０００７】
上述した課題を解決し、目的を達成するために、本開示の画像認識システムは、認識対象の画像データから検出対象を検出する画像認識システムであって、文章特徴抽出装置と、画像認識装置と、を備える。文章特徴抽出装置は、検出対象を指定するための自然言語の語句を取得する文章取得部と、自然言語の語句から画像の特徴量と比較可能な特徴量である文章特徴を抽出する文章特徴抽出部と、文章特徴を送信する文章特徴送信部と、を有する。画像認識装置は、文章特徴抽出装置から文章特徴を受信する文章特徴受信部と、文

章特徴受信部が受信した文章特徴を記憶する文章特徴記憶部と、画像データを取得する画像取得部と、画像データから画像内で検出対象が存在する領域の候補である候補領域、および、候補領域内の画像の特徴量である画像特徴を抽出する画像解析部と、文章特徴記憶部に記憶された文章特徴と、候補領域に対応する画像特徴とを比較して、比較結果を出力する特徴比較部と、比較結果に基づいて、検出対象の検出結果を生成する検出結果生成部と、を有する。画像認識システムは、文章特徴抽出部の処理を、画像認識装置とは別の装置である文章特徴抽出装置で実行することを特徴とする。

【発明の効果】

【０００８】

本開示によれば、画像認識装置の計算能力に制限がある場合であっても、学習済モデルを用いて、自然言語で入力される語句で指定された検出対象を画像データから検出する推論処理を実装することが可能な画像認識システムを得ることができるという効果を奏する。

【図面の簡単な説明】

【０００９】

【図１】実施の形態１にかかる画像認識システムの構成図

【図２】図１に示す画像認識システムが用いるＤＮＮモデルの構成例を示す図

【図３】図１に示す画像認識システムの動作例を説明するためのフローチャート

【図４】実施の形態２にかかる画像認識システムの構成図

【図５】図４に示す画像認識システムの動作例を説明するためのフローチャート

【図６】実施の形態１，２にかかる画像認識システムの機能を実現するための専用のハードウェアを示す図

【図７】実施の形態１，２にかかる画像認識システムの機能を実現するための制御回路の構成を示す図

【発明を実施するための形態】

【００１０】

以下に、本開示の実施の形態にかかる画像認識システム、画像認識装置、文章特徴抽出装置、プログラムおよび画像認識方法を図面に基づいて詳細に説明する。

【００１１】

実施の形態１．

図１は、実施の形態１にかかる画像認識システム１０の構成図である。画像認識システム１０は、画像認識装置１０００と、文章特徴抽出装置２０００と、結果表示部３０００とを有する。画像認識システム１０は、学習済モデルを用いて推論処理を行うことによって、自然言語で入力される語句で指定された検出対象を画像データから検出する機能を有する。学習済モデルは、例えば、Grounding DINOのような技術を用いて生成されたものであってよい。画像認識システム１０において、推論処理は、画像認識装置１０００と文章特徴抽出装置２０００とで分担して実行される。

【００１２】

具体的には、以下の実施の形態では、推論処理のうち負荷の高い処理、換言すれば計算量の多い処理として文章の特徴を抽出する処理を、文章特徴抽出装置２０００で実行し、その他の処理を画像認識装置１０００で実行する。ここでは、画像認識装置１０００と文章特徴抽出装置２０００とで分担する処理の割り当てを人間が予め決めることとするが、他の決め方であってもよい。

【００１３】

なお、本開示の一例として、文章特徴抽出装置２０００が複数の装置で構成されていたり、文章特徴抽出装置２０００が複数の演算装置を備えていたりする場合であっても、これらに含まれる複数の演算装置を１つのまとまりとして扱うこととする。同様に、画像認識装置１０００が複数の装置で構成されていたり、画像認識装置１０００が複数の演算装置を備えていたりする場合であっても、これらに含まれる複数の演算装置を１つのまとまりとして扱うこととする。つまり、画像認識装置１０００及び文章特徴抽出装置２０００のそれぞれの演算装置の構成によらず、上述の推論処理を実行するまとまりとして扱うこ

ととする。これにより、例えば、画像認識装置１０００が１つの演算装置を有する１台の撮像装置であって、文章特徴抽出装置２０００が複数の演算装置を有する１以上のサーバ装置であっても、画像認識装置１０００及び文章特徴抽出装置２０００をそれぞれ１つのまとまりとして演算処理能力の比較をすることができる。そして、本開示の例では、画像認識装置１０００よりも文章特徴抽出装置２０００のほうが演算処理能力が高いものとする。

【００１４】

画像認識装置１０００は、画像取得部１１００と、画像解析部１２００と、文章特徴受信部１３００と、特徴比較部１４００と、検出結果生成部１５００と、検出結果送信部１６００とを有する。画像認識装置１０００は、画像取得部１１００で取得した画像に指定された物体が映っているかどうか、また、映っている場合にはその物体の画像内に占める領域を認識し、認識結果を結果表示部３０００へ送信する。画像認識装置１０００は、例えば、画像認識のための処理装置または処理ソフトウェアを実行できるマイコンを内蔵した監視カメラ、デジタルカメラまたは監視カメラを接続した計算機などである。画像認識装置１０００は、上記の例に限定されない。また、本実施の形態では簡単のため、画像認識システム１０は、１台の画像認識装置１０００を有することとしたが、実施に当たっては、画像認識システム１０は、複数台の画像認識装置１０００を有してもよい。

【００１５】

文章特徴抽出装置２０００は、文章取得部２１００と、文章特徴抽出部２２００と、文章特徴送信部２３００とを有する。文章特徴抽出装置２０００は、画像認識装置１０００とは異なる装置である。文章特徴抽出装置２０００は、画像認識システム１０のユーザが、画像認識装置１０００に検出させたい物体カテゴリを、自然言語で入力される語句によって指定すると、入力された自然言語の語句から文章特徴を抽出し、抽出した文章特徴を画像認識装置１０００に送信する。ここで、文章特徴は、画像の特徴量と比較可能な特徴量である。

【００１６】

結果表示部３０００は、画像認識装置１０００が画像から物体検出を行った結果を表示するための装置である。結果表示部３０００は、表示機能を有していればよく、具体的な構成については特に制限はない。

【００１７】

以下、画像認識システム１０の各部の詳細について説明する。画像取得部１１００は、処理対象の画像データを取得し、取得した画像データを画像解析部１２００に出力する。画像取得部１１００は、例えば、ＣＣＤ（Charge Coupled Devices）、ＣＭＯＳ（Complementary Metal Oxide Semiconductor）などのイメージセンサ、或いは、計算機に接続されたデジタルカメラ、監視カメラなどの撮像装置である。

【００１８】

画像解析部１２００は、画像取得部１１００によって取得された画像データ内で物体が映っていると考えられる領域である候補領域を複数抽出し、抽出した複数の候補領域のそれぞれに含まれる特徴量である画像特徴を抽出する。画像解析部１２００は、抽出した複数の候補領域と、各候補領域に対応する画像特徴とを、特徴比較部１４００に出力する。画像解析部１２００は、事前に学習済のＤＮＮモデルのうちの一部であって、出力が、自然言語の入力に依存せず、画像データの入力にのみ依存する部分である第１の部分モデルを用いて、候補領域および画像特徴の抽出を行う。

【００１９】

画像解析部１２００は、候補領域の数と同数の画像特徴を抽出することができる。例えば、画像解析部１２００は、Ｎ個の候補領域を抽出した場合、Ｎ個の画像特徴を抽出することができる。

【００２０】

ここで、候補領域を示す情報の表現形式は、画像中で候補領域の範囲を表すことができれば、特に制限はない。例えば、候補領域の形状を幾何学形状として、幾何学形状の位置

を指定することで候補領域を表すことができる。幾何学形状は、例えば、矩形のような多角形であってもよいし、円であってもよい。候補領域の形状が多角形である場合、多角形の頂点の座標、辺の長さ、重心の座標などを用いて候補領域の位置を表すことができる。例えば、候補領域の形状が矩形である場合、矩形の４つの頂点座標で候補領域を表してもよいし、矩形の４つの頂点のうち対角線上に位置する２つの頂点の座標の組み合わせによって候補領域を表してもよいし、矩形の４つの頂点のうちの１つの座標と矩形の各辺の長さの数値とによって候補領域を表してもよいし、矩形の重心座標と矩形の各辺の長さの数値とによって候補領域を表してもよい。頂点座標、重心座標、各辺の長さの数値は、画像サイズと実座標との組み合わせであってもよいし、画像サイズに対する相対値であってもよい。

【００２１】

文章特徴受信部１３００は、文章特徴抽出装置２０００から送信された文章特徴を受信し、受信した文章特徴を特徴比較部１４００へ出力する。文章特徴は、指定された物体カテゴリの数だけあり、指定された物体カテゴリがＭ個ある場合には、文章特徴もＭ個受信される。文章特徴がＱ次元ベクトルである場合には、文章特徴受信部１３００は、Ｍ個のＱ次元ベクトルのデータを受信することになる。

【００２２】

特徴比較部１４００は、画像解析部１２００が出力する各候補領域に対応するＮ個の画像特徴と、文章特徴受信部１３００が出力するＭ個の文章特徴とを比較して類似度計算を行い、各画像特徴がＭ個の文章特徴のうちのどれに最も類似しているかを算出する。特徴比較部１４００は、類似度の算出結果を比較結果として検出結果生成部１５００へ出力する。特徴比較部１４００は、事前に学習済のＤＮＮモデルのうちの一部であって、画像解析部１２００が実行する第１の部分モデルの出力と、後述する文章特徴抽出部２２００が実行する第２の部分モデルの出力とを入力として受け付けて、類似度を比較結果として出力する第３の部分モデルを実行することによって、上述の処理を行う。

【００２３】

検出結果生成部１５００は、特徴比較部１４００が出力する比較結果に基づいて、各候補領域が最も類似している文章特徴に対応する物体カテゴリを特定し、特定した物体カテゴリを検出結果として、検出結果送信部１６００に出力する。また、物体カテゴリと候補領域の重複がある場合には、冗長な結果を検出結果から取り除く。

【００２４】

検出結果送信部１６００は、検出結果生成部１５００が出力する検出結果を、結果表示部３０００へ出力する。

【００２５】

結果表示部３０００は、検出結果送信部１６００から受信した検出結果に基づいて、画像認識システム１０のユーザに対して、所望の方法で検出結果の表示を行う。検出結果の表示方法については、特に制限はない。

【００２６】

文章取得部２１００は、画像認識システム１０のユーザが画像認識装置１０００に検出させたい物体カテゴリの自然言語の語句による入力を受け付け、受け付けた語句を文章特徴抽出部２２００に出力する。文章取得部２１００は、例えば、キーボード、タッチセンサなどの入力手段を用いた文字入力、マイクを用いた音声認識、予め作成された語句リストのデータを記憶装置から読み出すなどの方法で、自然言語の語句の入力を受け付けることができる。また、入力を受ける語句は複数あってよく、各語句は物体カテゴリに対応する。

【００２７】

文章特徴抽出部２２００は、文章取得部２１００で取得された自然言語の語句から文章特徴を抽出し、抽出した文章特徴を、文章特徴送信部２３００へ出力する。文章特徴抽出部２２００は、入力された自然言語の語句が複数ある場合には、それぞれの語句について文章特徴の抽出を行う。語句がＭ個ある場合、文章特徴抽出部２２００は、Ｍ個の文章特

徴の抽出を行う。文章特徴抽出部２２００は、事前に学習済のＤＮＮモデルのうちの一部であって、出力が、画像データの入力に依存せず、自然言語の語句の入力にのみ依存する第２の部分モデルを実行することによって、上述の処理を行う。

【００２８】

文章特徴送信部２３００は、文章特徴抽出部２２００が出力する文章特徴を、画像認識装置１０００の文章特徴受信部１３００へ送信する。文章特徴送信部２３００は、文章特徴が複数個ある場合には、それらの文章特徴を全て送信する。

【００２９】

図２は、図１に示す画像認識システム１０が用いるＤＮＮモデル＃０の構成例を示す図である。ＤＮＮモデル＃０は、入力を画像データおよび自然言語の語句とし、画像内から自然言語の語句に対応する物体を検出するように学習された学習済モデルである。このＤＮＮモデル＃０は、３つの部分モデルから構成されている。ＤＮＮモデル＃１は、出力が、画像データの入力にのみ依存し、自然言語の語句の入力に依存せず、独立して動作させることが可能な第１の部分モデルである。ＤＮＮモデル＃１は、画像データを入力として、画像内で検出対象が存在する領域の候補である候補領域と、候補領域毎の画像特徴とを出力する。ＤＮＮモデル＃１は、画像解析部１２００の処理に対応する。ＤＮＮモデル＃２は、出力が、自然言語の語句の入力にのみ依存し、画像データの入力に依存せず、独立して動作させることが可能な第２の部分モデルである。ＤＮＮモデル＃２は、自然言語の語句を入力として、語句毎の文章特徴を出力する。自然言語の語句は、画像認識装置１０００で検出させたい物体カテゴリを示すものであり、図２の例では、「男性」、「女性」、「杖を持った人」、「自転車に乗った人」といった、人物を検出するための語句が例示されている。ＤＮＮモデル＃２は、文章特徴抽出部２２００の処理に対応する。ＤＮＮモデル＃３は、ＤＮＮモデル＃０のうち、ＤＮＮモデル＃１およびＤＮＮモデル＃２の部分を除いた残りであり、第３の部分モデルである。ＤＮＮモデル＃３は、ＤＮＮモデル＃１の出力およびＤＮＮモデル＃２の出力を入力とし、文章特徴量と画像特徴量との類似度を計算する。ＤＮＮモデル＃３は、特徴比較部１４００の処理に対応する。

【００３０】

ここで、画像認識システム１０において基本的な処理の流れの説明をする前に、必要な初期処理について説明する。

【００３１】

まず、事前に、ＤＮＮモデル＃０を準備する。ＤＮＮモデル＃０は、十分に巨大なデータセットを用いて学習されている。画像内の領域の表現形式については、画像解析部１２００の説明において上述したものに従う。

【００３２】

続いて、ＤＮＮモデル＃１、ＤＮＮモデル＃２、ＤＮＮモデル＃３と同じ構造を持つＤＮＮモデルをそれぞれ定義し、学習済のＤＮＮモデル＃０から重み係数をそれぞれのＤＮＮモデルの対応する層にコピーしたうえで、ＤＮＮモデル＃１を画像解析部１２００に、ＤＮＮモデル＃２を文章特徴抽出部２２００に、ＤＮＮモデル＃３を特徴比較部１４００に、それぞれ配置する。各ＤＮＮモデルは、それぞれ画像認識装置１０００、文章特徴抽出装置２０００で計算可能な形で定義されているものとする。

【００３３】

以上が初期処理の内容である。続いて、画像認識システム１０での処理内容について説明する。図３は、図１に示す画像認識システム１０の動作例を説明するためのフローチャートである。画像認識装置１０００は、画像取得部１１００が画像データを取得するたびに以下の処理を行う。まず、画像取得部１１００は、周期的に画像データを取得する（ステップＳ２１１０）。取得周期は、画像認識システム１０の用途に応じて予め定められる。

【００３４】

続いて画像解析部１２００は、取得された画像データに対して、ＤＮＮモデル＃１を用いて、複数の候補領域の抽出と、各候補領域内の画像特徴の抽出とを行う（ステップＳ２１２０）。ここで得られた候補領域の数をＮ個とする。候補領域および画像特徴が得られ

たら、特徴比較部１４００は、候補領域が複数ある場合には、候補領域の中から１つを選択して、候補領域毎の特徴量比較処理を開始し（ステップＳ２１３０）、文章特徴受信部１３００が文章特徴を受信するのを待機する。

【００３５】

一方、文章特徴抽出装置２０００では、上述のステップＳ２１１０からステップＳ２１３０の処理と並行して、以下の処理が行われる。文章取得部２１００は、画像認識システム１０のユーザが入力する自然言語の語句を取得する文章取得処理を行い（ステップＳ２２１０）、取得した自然言語の語句を、文章特徴抽出部２２００に出力する。自然言語の語句は、画像認識装置１０００で検出させたい物体カテゴリを示すものである。

【００３６】

文章特徴抽出部２２００は、文章取得部２１００が出力する自然言語の語句から文章特徴を抽出する文章特徴抽出処理を開始する（ステップＳ２２２０）。文章特徴抽出部２２００は、ＤＮＮモデル＃２を用いて、与えられた自然言語の語句から文章特徴の抽出を行う（ステップＳ２２３０）。文章特徴抽出部２２００は、文章取得部２１００が取得した全ての語句を処理済みであるか否かを判断する（ステップＳ２２４０）。未だ処理済みでない語句がある場合（ステップＳ２２４０：Ｎｏ）、文章特徴抽出部２２００は、次の語句について、ステップＳ２２３０の処理を繰返す。全ての語句を処理済みである場合（ステップＳ２２４０：Ｙｅｓ）、文章特徴抽出部２２００は、文章特徴抽出処理を終了する（ステップＳ２２５０）。文章特徴抽出部２２００は、自然言語の語句の個数をＭ個とした場合、ステップＳ２２２０からステップＳ２２５０の処理をＭ回繰り返す。

【００３７】

各自然言語の語句に対応するＭ個の文章特徴が得られたら、文章特徴送信部２３００は、画像認識装置１０００が有する文章特徴受信部１３００へ、文章特徴を自然言語の語句と対応づけて送信する（ステップＳ２２６０）。

【００３８】

文章特徴受信部１３００が自然言語の語句および文章特徴を受信すると、特徴比較部１４００は、受信した自然言語の語句の中から１つを選択して、語句毎の特徴量比較処理を開始する（ステップＳ２１４０）。特徴比較部１４００は、選択された候補領域の画像特徴と、選択された語句の文章特徴との特徴量比較を行う（ステップＳ２１５０）。具体的には、特徴比較部１４００は、ＤＮＮモデル＃３を用いて、画像特徴と文章特徴との類似度を計算する。

【００３９】

特徴比較部１４００は、対象の候補領域について、全ての語句と特徴量を比較済みであるか否かを判断する（ステップＳ２１６０）。未だ比較済みでない語句がある場合（ステップＳ２１６０：Ｎｏ）、特徴比較部１４００は、次の語句を選択して、ステップＳ２１５０から処理を繰返す。全ての語句と比較済みである場合（ステップＳ２１６０：Ｙｅｓ）、特徴比較部１４００は、対象の候補領域について、語句毎の特徴量比較処理を終了し（ステップＳ２１７０）、全ての候補領域について特徴量を比較済みであるか否かを判断する（ステップＳ２１８０）。未だ比較済みでない候補領域がある場合（ステップＳ２１８０：Ｎｏ）、特徴比較部１４００は、次の候補領域を選択して、ステップＳ２１４０から処理を繰返す。全ての候補領域について特徴量を比較済みである場合（ステップＳ２１８０：Ｙｅｓ）、特徴比較部１４００は、候補領域毎の特徴量比較処理を終了する（ステップＳ２１９０）。ここで、Ｍ個の語句の文章特徴とＮ個の候補領域の画像特徴との全ての組み合わせについて類似度の計算が終わると、特徴比較部１４００は、候補領域毎に、類似度が最も高い語句を記憶する。

【００４０】

検出結果生成部１５００は、特徴比較部１４００の比較結果を取得すると、各候補領域の情報が画像データの幅および高さに対して相対値であれば、候補領域の情報を元の画像座標基準値へ換算し、候補領域同士が重なっているものがあれば、領域の重なりの大きさと、対応づけられた文章特徴同士とを比較し、重なりが十分に大きく、且つ、対応づけら

れた文章特徴が同一のものが複数あれば、重複している候補領域であると判断して、重複したもののうち１つを残して不要な候補領域を除去し（ステップＳ２３１０）、検出結果を生成し、生成した検出結果を検出結果送信部１６００へ出力する。

【００４１】

検出結果送信部１６００は、検出結果生成部１５００が出力する検出結果を結果表示部３０００へ送信する（ステップＳ２３２０）。検出結果を送信すると、画像認識装置１０００は、次の画像データの取得を待機し、再びステップＳ２１１０から処理を開始する。

【００４２】

以上が実施の形態１にかかる画像認識システム１０における物体検出を行う処理の流れである。通常、ＤＮＮモデル＃０の処理は一体の計算装置を用いて行われるところ、本実施の形態においては、画像データの特徴抽出結果に依存しない、文章特徴の抽出処理を分離して、他の計算機で実施することで、計算能力に制限のある機器、例えば、監視カメラなどの組込機器での物体検出を容易にしている。ＤＮＮモデル＃０は、画像データと、画像データから検出する検出対象となる物体のカテゴリを指定する自然言語の語句とを入力として受け付け、検出結果を出力する学習済モデルであり、十分に巨大なデータセットを用いて学習されている。このため、画像認識システム１０のユーザは、検出対象となる物体カテゴリを変更したい場合には、文章取得部２１００へ与える自然言語の語句を変更するだけで、ＤＮＮモデルの再学習なしで検出対象となる物体カテゴリを変更することができる。

【００４３】

なお、図３のステップＳ２１３０からステップＳ２１９０のループでは、１つの画像特徴に対して複数の文章特徴毎の比較処理を行うループを、画像特徴の数だけ繰り返すという順番で行っているが、ループの順序を逆にしてもよい。すなわち、１つの文章特徴に対して複数の画像特徴毎の比較処理を行うループを、文章特徴の数だけ繰り返すという順番で行っても同じ結果が得られる。

【００４４】

また、特徴比較部１４００の処理は、ＤＮＮモデル＃３を用いるとしたが、ＤＮＮモデル＃０の構造によっては、この部分は必ずしもＤＮＮモデルでなくても構わない。例えば、特徴比較部１４００は、画像特徴と文章特徴とのコサイン類似度を計算するような処理部であってもよいし、或いは、画像特徴と文章特徴とのユークリッド距離を計算するような処理部であってもよい。ただし、ＤＮＮモデル＃０を事前に学習するときに、上記した処理部を適用して十分に学習されていることが前提となる。

【００４５】

また、説明のため、図１では結果表示部３０００と文章特徴抽出装置２０００とを別体として示したが、文章特徴抽出装置２０００が結果表示部３０００の機能を兼ねてもよい。

【００４６】

以上説明したように、実施の形態１によれば、画像認識装置１０００と文章特徴抽出装置２０００とを備え、認識対象の画像データから検出対象を検出する画像認識システム１０を提供することができる。文章特徴抽出装置２０００は、検出対象を指定するための自然言語の語句を取得する文章取得部２１００と、自然言語の語句から画像の特徴量と比較可能な特徴量である文章特徴を抽出する文章特徴抽出部２２００と、文章特徴を送信する文章特徴送信部２３００と、を有する。画像認識装置１０００は、文章特徴抽出装置２０００から文章特徴を受信する文章特徴受信部１３００と、画像データを取得する画像取得部１１００と、画像データから画像内で検出対象が存在する領域の候補である候補領域、および、候補領域内の画像の特徴量である画像特徴を抽出する画像解析部１２００と、文章特徴と、候補領域に対応する画像特徴とを比較して、比較結果を出力する特徴比較部１４００と、比較結果に基づいて、検出対象の検出結果を生成する検出結果生成部１５００と、を有する。画像認識システム１０は、文章特徴抽出部２２００の処理を、画像認識装置１０００とは別の装置である文章特徴抽出装置２０００で実行することを特徴とする。これにより、自然言語の語句から文章特徴を抽出する負荷の高い処理を、画像認識装置１

０００とは異なる装置である文章特徴抽出装置２０００で実行させることができるため、画像認識装置１０００の計算能力に制限がある場合であっても、学習済モデルを用いて、自然言語で入力される語句で指定された検出対象を画像データから検出する推論処理を実装することが可能になる。

【００４７】

なお、画像解析部１２００は、画像データから自然言語の語句で指定された検出対象を検出するための学習済モデルであるＤＮＮモデル＃０のうちの一部であって、出力が自然言語の入力に依存しない第１の部分モデルであるＤＮＮモデル＃１を実行することによって実現される。ＤＮＮモデル＃０は、例えば、生成ＡＩ（Artificial Intelligence）である。画像データおよび自然言語の語句のようにマルチモーダルなデータを入力とする生成ＡＩは、計算量が非常に多く、必要な計算能力が非常に高くなるため、監視カメラのような組込機器で動作させる場合には、計算能力が足りず、実装が困難な場合がある。本実施の形態の技術を用いることで、監視カメラのような組込機器において、計算能力が足りず、マルチモーダルなデータをまとめて処理することが困難な場合であっても、処理負荷の高い一部の処理を外部装置で実行することが可能になる。

【００４８】

また、画像認識システム１０において、文章特徴抽出部２２００は、画像データから自然言語の語句で指定された検出対象を検出するための学習済モデルであるＤＮＮモデル＃０の一部であって、出力が画像データの入力に依存しない第２の部分モデルであるＤＮＮモデル＃２を実行することによって実現されてもよい。

【００４９】

また、画像認識システム１０において、特徴比較部１４００は、画像データから自然言語の語句で指定された検出対象を検出するための学習済モデルであるＤＮＮモデル＃０の一部であって、自然言語の入力に依存しない第１の部分モデルであるＤＮＮモデル＃１の出力と、画像データの入力に依存しない第２の部分モデルであるＤＮＮモデル＃２の出力とを入力として受け付けて、文章特徴と画像特徴との比較結果を出力する第３の部分モデルであるＤＮＮモデル＃３を実行することによって実現されてもよい。

【００５０】

実施の形態１によれば、認識対象の画像データから検出対象を検出する画像認識装置１０００を提供することもできる。画像認識装置１０００は、画像の特徴量と比較可能な特徴量であって検出対象を指定するための自然言語の語句から抽出された文章特徴を画像認識装置１０００の外部から受信する文章特徴受信部１３００と、画像データを取得する画像取得部１１００と、画像データから画像内で検出対象が存在する領域の候補である候補領域、および、候補領域内の画像の特徴量である画像特徴を抽出する画像解析部１２００と、文章特徴と、候補領域に対応する画像特徴とを比較して、比較結果を出力する特徴比較部１４００と、比較結果に基づいて、検出対象の検出結果を生成する検出結果生成部１５００と、を備えることを特徴とする。

【００５１】

また、画像認識装置１０００において、画像解析部１２００は、画像データから自然言語の語句で指定された検出対象を検出するための学習済モデルであるＤＮＮモデル＃０のうちの一部であって、出力が自然言語の入力に依存しない第１の部分モデルであるＤＮＮモデル＃１を実行することによって実現されてもよい。また、画像認識装置１０００において、特徴比較部１４００は、上記の学習済モデルであるＤＮＮモデル＃０の一部であって、第１の部分モデルであるＤＮＮモデル＃１の出力と、上記の学習済モデルであるＤＮＮモデル＃０の一部であって画像データの入力に依存しない第２の部分モデルであるＤＮＮモデル＃２の出力とを入力として受け付けて、比較結果を出力する第３の部分モデルであるＤＮＮモデル＃３を実行することによって実現されてもよい。

【００５２】

また、実施の形態１によれば、文章特徴抽出装置２０００を提供することもできる。文章特徴抽出装置２０００は、画像データから検出する検出対象を指定するための自然言語

の語句を取得する文章取得部２１００と、自然言語の語句から画像の特徴量と比較可能な特徴量である文章特徴を抽出する文章特徴抽出部２２００と、文章特徴を、画像データから検出対象を検出する画像認識装置１０００に送信する文章特徴送信部２３００と、を備えることを特徴とする。

【００５３】

文章特徴抽出装置２０００において、文章特徴抽出部２２００は、画像データから自然言語の語句で指定された検出対象を検出するための学習済モデルであるＤＮＮモデル＃０の一部であって、出力が画像データの入力に依存しない第２の部分モデルであるＤＮＮモデル＃２を実行することによって実現することができる。

【００５４】

また、ハードウェア構成については後述するが、実施の形態１によれば、コンピュータを、認識対象の画像データから検出対象を検出する画像認識装置１０００として機能させるためのプログラムを提供することもできる。このプログラムは、画像の特徴量と比較可能な特徴量であって、当該コンピュータとは異なるコンピュータにおいて検出対象を指定するための自然言語の語句から抽出された文章特徴を受信するステップと、画像データを取得するステップと、画像データから画像内で検出対象が存在する領域の候補である候補領域、および、候補領域内の画像の特徴量である画像特徴を抽出するステップと、文章特徴と、候補領域に対応する画像特徴とを比較して、比較結果を出力するステップと、比較結果に基づいて、検出対象の検出結果を生成するステップと、を画像認識装置１０００として機能させるコンピュータに実行させることができる。

【００５５】

また、実施の形態１によれば、コンピュータを、文章特徴抽出装置２０００として機能させるためのプログラムを提供することもできる。このプログラムは、画像データから検出する検出対象を指定するための自然言語の語句を取得するステップと、自然言語の語句から画像の特徴量と比較可能な特徴量である文章特徴を抽出するステップと、文章特徴を、文章特徴抽出装置２０００として機能させるコンピュータとは異なるコンピュータであって、画像データから検出対象を検出する画像認識装置１０００として機能するコンピュータに送信するステップと、を文章特徴抽出装置２０００として機能させるコンピュータに実行させることができる。

【００５６】

また、実施の形態１によれば、認識対象の画像データから検出対象を検出する画像認識方法を提供することができる。この画像認識方法は、検出対象を指定するための自然言語の語句から画像の特徴量と比較可能な特徴量である文章特徴を抽出するステップと、画像データから画像内で検出対象が存在する領域の候補である候補領域、および、候補領域内の画像の特徴量である画像特徴を抽出するステップと、文章特徴と、候補領域に対応する画像特徴とを比較して、比較結果を生成するステップと、比較結果に基づいて、検出対象の検出結果を生成するステップと、を含む。また、この画像認識方法は、文章特徴を抽出するステップの処理と、画像特徴を抽出するステップの処理とを異なる装置で実行することを特徴とする。例えば、画像認識装置１０００および文章特徴抽出装置２０００を備える画像認識システム１０により実行され、文章特徴抽出装置２０００が文章特徴を抽出するステップを実行し、画像認識装置１０００が候補領域および画像特徴を抽出するステップと、比較結果を生成するステップと、検出結果を生成するステップとを実行することができる。また、画像認識方法は、自然言語の語句を取得するステップと、抽出した文章特徴を送信するステップと、文章特徴を受信するステップと、をさらに含んでもよい。この場合、文章特徴抽出装置２０００が、自然言語の語句を取得するステップと、抽出した文章特徴を送信するステップと、を実行し、画像認識装置１０００が、文章特徴を受信するステップを実行し、受信した文章特徴を用いて比較結果を生成するステップを実行する。上述の画像認識方法において、記載した各ステップの処理は、必ずしも、記載した順で実行される必要はない。画像認識方法は、同時並行で実行される複数のステップを含んでいてもよい。また、画像認識方法において、記載した各ステップの処理は、必ずしも連続し

て実行される一連の処理でなくてもよく、異なるタイミングで実行される複数の処理を含んでいてもよい。

【００５７】

実施の形態２．

図４は、実施の形態２にかかる画像認識システム２０の構成図である。画像認識システム２０は、画像認識装置１０５０と、文章特徴抽出装置２０００と、結果表示部３０００とを有する。画像認識システム２０は、画像認識システム１０の画像認識装置１０００の代わりに画像認識装置１０５０を有する。以下、実施の形態１と異なる点を主に説明し、実施の形態１と同様の部分については詳細な説明を省略する。

【００５８】

画像認識システム２０は、文章特徴の抽出処理を、検出対象となる物体カテゴリの変更または追加があったときのみ実行する。

【００５９】

画像認識装置１０５０は、画像取得部１１００と、画像解析部１２００と、文章特徴受信部１３００と、特徴比較部１４００と、検出結果生成部１５００と、検出結果送信部１６００とを有する。画像認識装置１０５０は、画像認識装置１０００の構成に加えて、文章特徴記憶部１３５０を有する。

【００６０】

文章特徴記憶部１３５０は、文章特徴受信部１３００が文章特徴抽出装置２０００から受信する文章特徴を記憶する。文章特徴記憶部１３５０は、複数の文章特徴を記憶することが可能であり、記憶した文章特徴を特徴比較部１４００に出力することができる。

【００６１】

図５は、図４に示す画像認識システム２０の動作例を説明するためのフローチャートである。図５は、図３のフローチャートにステップＳ２３３０を加えたものである。ステップＳ２３３０以外のステップは図３と同一であるため、各ステップの詳細については説明を省略し、以下、図３と異なる動作のみ説明する。

【００６２】

画像認識システム２０では、文章特徴の抽出処理であるステップＳ２２１０からステップＳ２２５０までの一連の処理は、画像特徴抽出処理と必ずしも並列に実行されず、基本的には、画像特徴抽出処理よりも早い任意のタイミングで独立して実行される。文章特徴送信部２３００が、ステップＳ２２６０で文章特徴を送信したのち、文章特徴受信部１３００は、受信した文章特徴を、文章特徴記憶部１３５０へ記憶させる（ステップＳ２３３０）。

【００６３】

文章特徴記憶部１３５０へ文章特徴が記憶された後、特徴比較部１４００は、ステップＳ２１５０で、文章特徴と画像特徴との比較処理を実行する。以降の処理は、図３と同様である。

【００６４】

また、任意のタイミングでステップＳ２２１０の文章取得処理が起こると、文章特徴抽出装置２０００は、再度、ステップＳ２２２０からステップＳ２２６０の処理を実行して、文章特徴受信部１３００は、文章特徴記憶部１３５０の内容を更新する。

【００６５】

以上、実施の形態２にかかる画像認識システム２０において、画像認識装置１０５０は、文章特徴受信部１３００が受信した文章特徴を記憶する文章特徴記憶部１３５０をさらに有し、特徴比較部１４００は、文章特徴記憶部１３５０に記憶された文章特徴と画像特徴とを比較することができる。このような構成を有することにより、画像データの取得よりも前に、文章特徴の抽出を実行して画像認識装置１０５０に記憶しておくことが可能になる。これにより、検出対象の変更がない場合には、文章特徴の抽出を実行する必要がなくなり、特徴比較部１４００は、文章特徴の受信を待つことなく、予め文章特徴記憶部１３５０に記憶された文章特徴を使用して特徴量比較処理を実行することが可能になる。し

たがって、画像認識装置１０５０のスループットが向上するとともに、画像認識システム２０全体としての計算負荷を低減することができる。

【００６６】

この場合、文章特徴抽出装置２０００は、文章取得部２１００が新たな自然言語の語句を取得する毎に、文章特徴抽出部２２００によって、新たな自然言語の語句から文章特徴を抽出し、抽出した文章特徴を文章特徴送信部２３００によって画像認識装置１０５０へ送信することができる。新たな自然言語の語句の取得は、例えば、画像認識処理の実行毎に行われなくてもよく、最初に検出対象が設定された後は、検出対象の変更、つまり、検出対象の物体カテゴリを追加したり、検出対象に設定した物体カテゴリを削除したりといった変更が行われるまでは、基本的には、文章特徴記憶部１３５０に記憶された文章特徴を用いて画像認識が実行される。物体カテゴリの追加が行われる場合、文章取得部２１００が、新たな自然言語の語句を取得するため、文章特徴抽出部２２００によって、新たに文章特徴の抽出が行われる。通常、ＤＮＮモデルを用いて、自然言語の語句および画像データを入力とした推論処理を行う場合、画像取得のたびにＤＮＮモデル全体を動作させる必要がある。この場合、検出対象の変更がなければ、実際には文章特徴抽出部２２００の処理内容および結果に変化はないため、計算効率が悪い。これに対して、実施の形態２の技術によれば、検出対象の変更が行われない限り、文章特徴の抽出処理が省略されるため、画像認識装置１０５０のスループットが向上するとともに、画像認識システム２０全体としての計算負荷を低減することができる。

【００６７】

ここで、画像認識システム１０，２０の機能を実現するためのハードウェア構成について説明する。文章取得部２１００、文章特徴抽出部２２００、画像取得部１１００、画像解析部１２００、特徴比較部１４００および検出結果生成部１５００は、処理回路により実現される。これらの処理回路は、専用のハードウェアにより実現されてもよいし、ＣＰＵ（Central Processing Unit）を用いた制御回路であってもよい。

【００６８】

上記の処理回路が、専用のハードウェアにより実現される場合、これらは、図６に示す処理回路９０により実現される。図６は、実施の形態１，２にかかる画像認識システム１０，２０の機能を実現するための専用のハードウェアを示す図である。処理回路９０は、単一回路、複合回路、プログラム化したプロセッサ、並列プログラム化したプロセッサ、ＡＳＩＣ（Application Specific Integrated Circuit）、ＦＰＧＡ（Field Programmable Gate Array）、またはこれらを組み合わせたものである。

【００６９】

上記の処理回路が、ＣＰＵを用いた制御回路で実現される場合、この制御回路は例えば図７に示す構成の制御回路９１である。図７は、実施の形態１，２にかかる画像認識システム１０，２０の機能を実現するための制御回路の構成を示す図である。図７に示すように、制御回路９１は、プロセッサ９２と、メモリ９３とを備える。プロセッサ９２は、ＣＰＵであり、処理装置、演算装置、マイクロプロセッサ、マイクロコンピュータ、ＤＳＰ（Digital Signal Processor）などとも呼ばれる。メモリ９３は、例えば、ＲＡＭ（Random Access Memory）、ＲＯＭ（Read Only Memory）、フラッシュメモリ、ＥＰＲＯＭ（Erasable Programmable ＲＯＭ）、ＥＥＰＲＯＭ（登録商標）（Electrically ＥＰＲＯＭ）などの不揮発性または揮発性の半導体メモリ、磁気ディスク、フレキシブルディスク、光ディスク、コンパクトディスク、ミニディスク、ＤＶＤ（Digital Versatile Disk）などである。

【００７０】

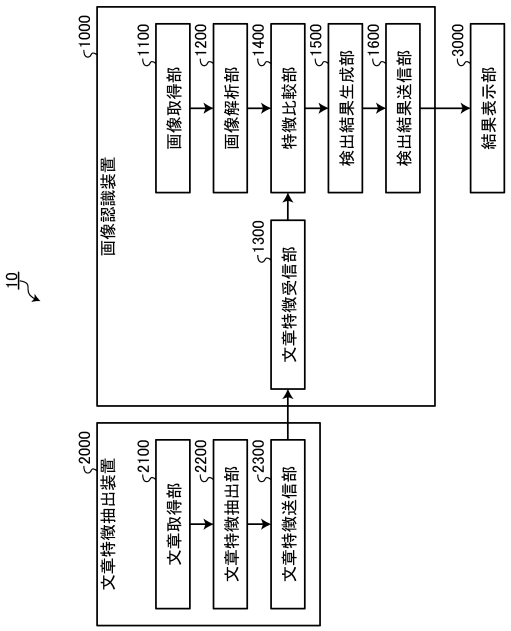
上記の処理回路が制御回路９１により実現される場合、プロセッサ９２がメモリ９３に記憶された、各構成要素の処理に対応するプログラムを読み出して実行することにより実現される。また、メモリ９３は、プロセッサ９２が実行する各処理における一時メモリとしても使用される。プロセッサ９２が実行するプログラムは、記憶媒体に記憶された状態で提供されてもよいし、インターネットのような通信路を介して提供されてもよい。

【００７１】
　また、文章特徴受信部１３００、検出結果送信部１６００および文章特徴送信部２３００の機能は、図示しない通信装置を用いて実現することができる。また、画像取得部１１００の機能は、上述の通り、イメージセンサや撮像装置を用いて実現することができる。文章特徴記憶部１３５０は、記憶装置を用いて実現することができる。
【００７２】
　以上の実施の形態に示した構成は、一例を示すものであり、別の公知の技術と組み合わせることも可能であるし、実施の形態同士を組み合わせることも可能であるし、要旨を逸脱しない範囲で、構成の一部を省略、変更することも可能である。
【００７３】
　例えば、図１および図４では、簡単のため、画像認識装置１０００，１０５０を１台、文章特徴抽出装置２０００を１台のみ示しているが、画像認識システム１０，２０は、複数の画像認識装置１０００，１０５０を備えていてもよい。文章特徴抽出装置２０００は、複数台の画像認識装置１０００，１０５０に文章特徴を送信することができる。また、必要とされる計算能力に応じて、画像認識システム１０，２０は、複数台の文章特徴抽出装置２０００を備えていてもよい。
【００７４】
　また、上記の実施の形態では、画像認識装置１０００，１０５０の画像解析部１２００が複数の候補領域を抽出した場合について説明したが、画像から抽出できる候補領域が１つの場合もある。この場合、図３および図５のステップＳ２１３０からステップＳ２１９０のループ処理が１回だけ実行されることとなる。
【００７５】
　なお、上記の実施の形態では、画像認識装置１０００，１０５０は、文章特徴抽出装置２０００から直接、文章特徴を受信することとしたが、かかる例に限定されない。画像認識装置１０００，１０５０以外で文章特徴抽出処理が行われればよく、例えば、文章特徴抽出装置２０００が抽出した文章特徴を、文章特徴抽出装置２０００以外のコンピュータで記憶しておき、画像認識装置１０００，１０５０は、文章特徴を記憶するコンピュータから文章特徴を受信してもよい。
【符号の説明】
【００７６】
　１０，２０　画像認識システム、９０　処理回路、９１　制御回路、９２　プロセッサ、９３　メモリ、１０００，１０５０　画像認識装置、１１００　画像取得部、１２００　画像解析部、１３００　文章特徴受信部、１３５０　文章特徴記憶部、１４００　特徴比較部、１５００　検出結果生成部、１６００　検出結果送信部、２０００　文章特徴抽出装置、２１００　文章取得部、２２００　文章特徴抽出部、２３００　文章特徴送信部、３０００　結果表示部。

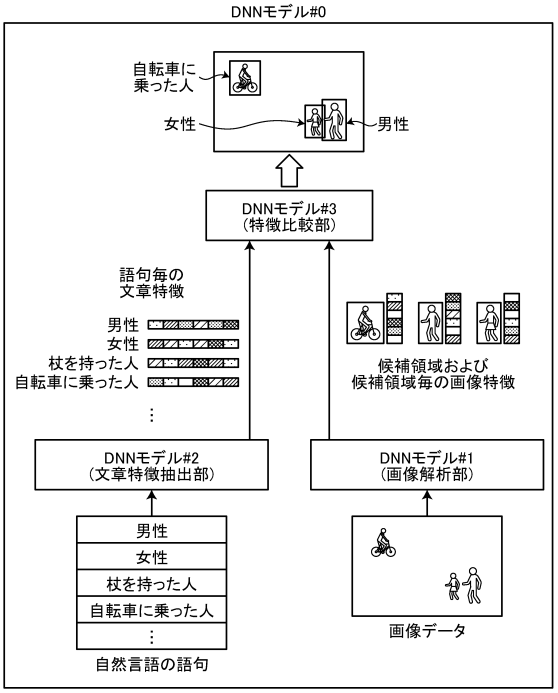
【要約】
画像認識システム（１０）は、文章特徴抽出装置（２０００）と、画像認識装置（１０００）と、を備える。文章特徴抽出装置（２０００）は、ユーザが入力した自然言語の語句を取得する文章取得部（２１００）と、自然言語の語句から文章特徴を抽出する文章特徴抽出部（２２００）と、文章特徴を送信する文章特徴送信部（２３００）と、を有する。画像認識装置（１０００）は、文章特徴抽出装置（２０００）から文章特徴を受信する文章特徴受信部（１３００）と、画像データを取得する画像取得部（１１００）と、画像データから候補領域、および、候補領域内の画像の特徴量である画像特徴を抽出する画像解析部（１２００）と、文章特徴と、候補領域に対応する画像特徴とを比較して、比較結果を出力する特徴比較部（１４００）と、比較結果に基づいて、検出対象の検出結果を生成する検出結果生成部（１５００）と、を有する。

【図面】

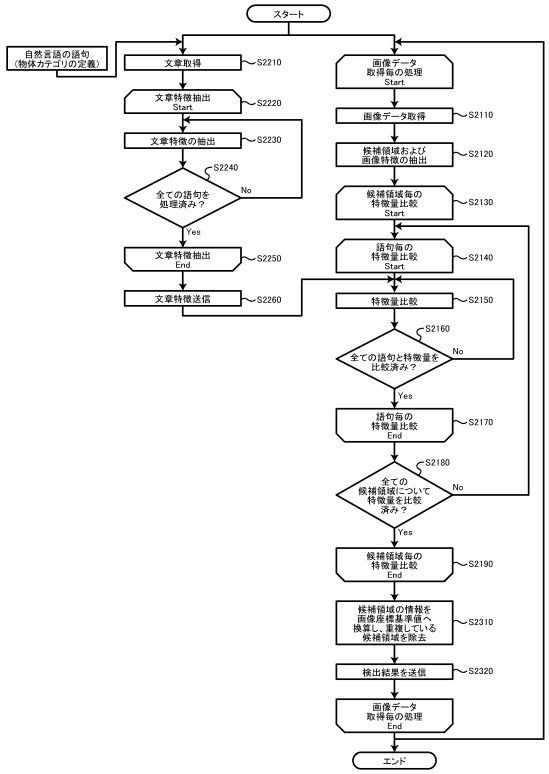
【図１】



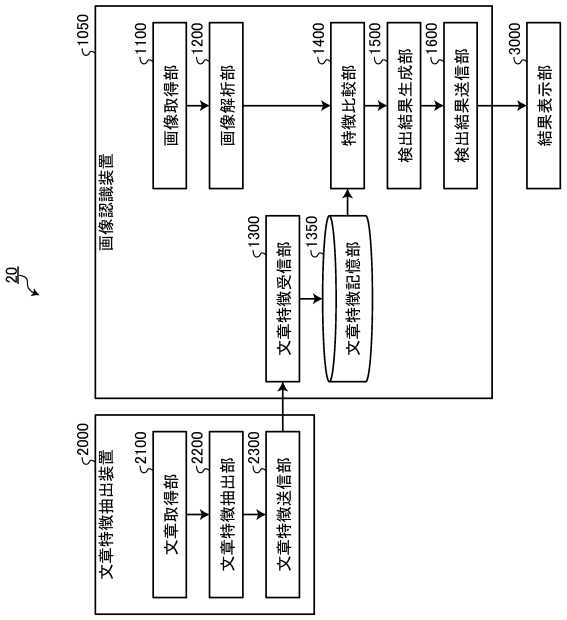
【図２】



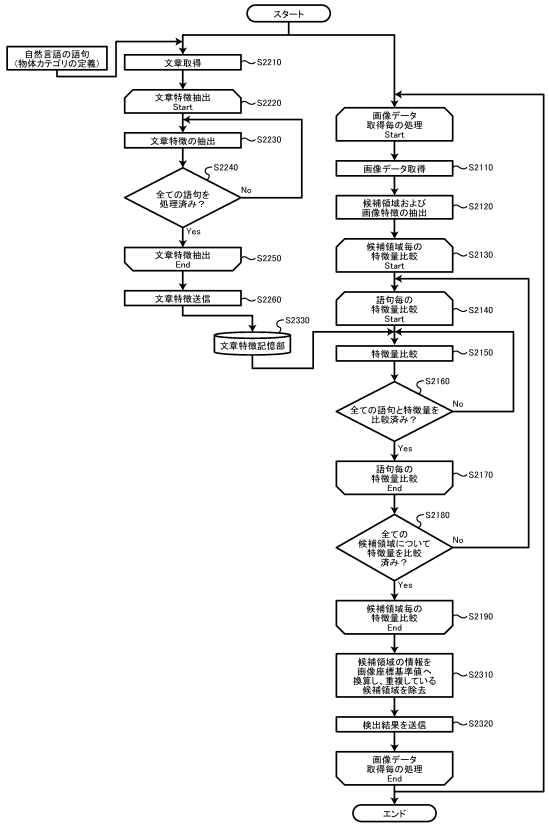
【図３】


スタート
自然言語の語句（物体カテゴリの定義）
文章取得
S2210
文章特徴抽出 Start
S2220
文章特徴の抽出
S2230
全ての語句を処理済み？
S2240
No
Yes
文章特徴抽出 End
S2250
文章特徴送信
S2260
画像データ取得毎の処理 Start
画像データ取得
S2110
候補領域および画像特徴の抽出
S2120
候補領域毎の特徴量比較 Start
S2130
語句毎の特徴量比較 Start
S2140
特徴量比較
S2150
全ての語句と特徴量を比較済み？
S2160
語句毎の特徴量比較 End
S2170
全ての候補領域について特徴量を比較済み？
S2180
候補領域毎の特徴量比較 End
S2190
候補領域の情報を画像座標基準値へ換算し、重複している候補領域を除去
S2310
検出結果を送信
S2320
画像データ取得毎の処理 End
エンド

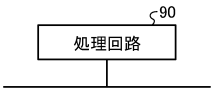
【図４】


20
画像認識装置
1050
画像取得部
1100
画像解析部
1200
特徴比較部
1400
検出結果生成部
1500
検出結果送信部
1600
結果表示部
3000
文章特徴受信部
1300
文章特徴記憶部
1350
文章特徴抽出装置
2000
文章取得部
2100
文章特徴抽出部
2200
文章特徴送信部
2300

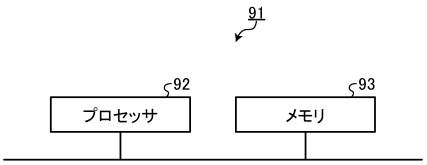
【図５】


スタート
自然言語の語句（物体カテゴリの定義）
文章取得
S2210
文章特徴抽出 Start
S2220
文章特徴の抽出
S2230
全ての語句を処理済み？
S2240
No
Yes
文章特徴抽出 End
S2250
文章特徴送信
S2260
文章特徴記憶部
S2330
画像データ取得毎の処理 Start
画像データ取得
S2110
候補領域および画像特徴の抽出
S2120
候補領域毎の特徴量比較 Start
S2130
語句毎の特徴量比較 Start
S2140
特徴量比較
S2150
全ての語句と特徴量を比較済み？
S2160
語句毎の特徴量比較 End
S2170
全ての候補領域について特徴量を比較済み？
S2180
候補領域毎の特徴量比較 End
S2190
候補領域の情報を画像座標基準値へ換算し、重複している候補領域を除去
S2310
検出結果を送信
S2320
画像データ取得毎の処理 End
エンド

【図６】


90
処理回路

【図７】



フロントページの続き

(56)参考文献 特開２０２２－１０６１４７（ＪＰ，Ａ）
Alee Radford et.al，Learning Transforable Visual Models From Natural Language Supervision，arXiv，米国，Cornell University，2021年02月26日，p.1-p.48，https://arxiv.org/pdf/2103.00020v1

(58)調査した分野 (Int.Cl.，ＤＢ名)
Ｇ０６Ｔ ７／００－７／９０