(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2019-96006
(P2019-96006A)
(43)公開日　令和1年6月20日(2019.6.20)

(51)Int.Cl.			F I			テーマコード（参考）
G06T	7/00	(2017.01)	G06T	7/00	350C	5L096
G06N	3/04	(2006.01)	G06N	3/04		
G06N	3/08	(2006.01)	G06N	3/08		

審査請求 未請求　請求項の数 31　ＯＬ　（全 40 頁）

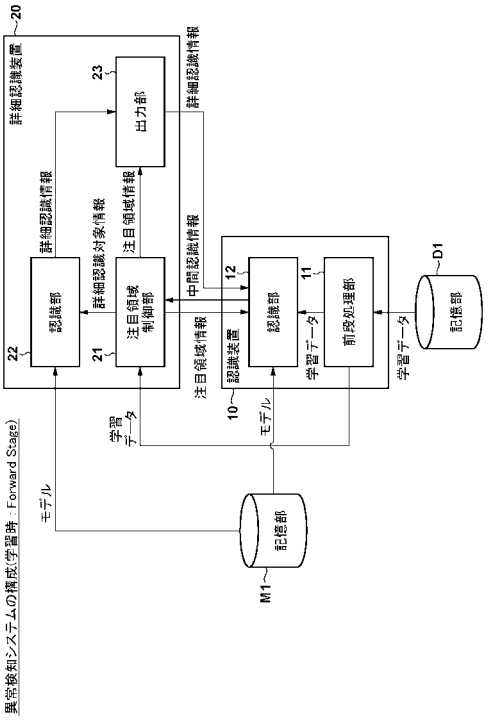
(21)出願番号　特願2017-223929 (P2017-223929)
(22)出願日　平成29年11月21日 (2017.11.21)

(71)出願人　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　100076428
弁理士　大塚　康徳
(74)代理人　100115071
弁理士　大塚　康弘
(74)代理人　100112508
弁理士　高柳　司郎
(74)代理人　100116894
弁理士　木村　秀二
(74)代理人　100130409
弁理士　下山　治
(74)代理人　100134175
弁理士　永川　行光

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)【要約】
【課題】　認識対象の認識精度を向上させるための技術を提供すること。
【解決手段】　認識対象の情報を含む認識対象データを入力として畳み込みニューラルネットワークの認識処理の結果を取得する。認識対象データおよび／または畳み込みニューラルネットワークの中間層出力に対する注目領域を設定する。注目領域中の認識対象データおよび／または中間層出力に対して詳細認識処理を行う。詳細認識処理の結果と中間層出力とを統合処理する。統合処理の結果を中間層出力として畳み込みニューラルネットワークに入力する。認識処理の結果を出力する。
【選択図】　図１



【特許請求の範囲】
【請求項１】
認識対象の情報を含む認識対象データを入力として畳み込みニューラルネットワークの認識処理の結果を取得する認識処理部と、
前記認識対象データおよび／または前記畳み込みニューラルネットワークの中間層出力に対する注目領域を設定する注目領域設定部と、
前記注目領域中の前記認識対象データおよび／または前記中間層出力に対して前記認識処理よりも詳細な詳細認識処理を行う詳細認識部と、
前記詳細認識処理の結果と、前記中間層出力と、を統合処理する統合処理部と、
前記統合処理の結果を前記中間層出力として前記畳み込みニューラルネットワークに入力する中間入力処理部と、
前記認識処理の結果を出力する出力部と
を有することを特徴とする情報処理装置。
【請求項２】
前記詳細認識部が有する認識器が前記畳み込みニューラルネットワークであ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】
前記詳細認識部は少なくとも二つの認識器を備えることを特徴とする請求項１又は２に記載の情報処理装置。
【請求項４】
前記注目領域設定部は、前記認識対象の一部に基づいて前記注目領域を設定し、
前記詳細認識部は、前記一部の種類ごとに認識器を備え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】
前記一部は少なくとも人体の部位に対応する情報を含むことを特徴とする請求項４に記載の情報処理装置。
【請求項６】
前記認識対象データは画像を含み、
前記統合処理部は、複数の前記一部の前記画像における位置に基づいて前記統合処理を行うことを特徴とする請求項４又は５に記載の情報処理装置。
【請求項７】
前記認識処理部は、前記認識対象データに含まれる物体または現象に関する異常な状態を検知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】
前記認識処理部および前記詳細認識部は、マルチタスク学習に基づいて前記認識処理を行うことを特徴とする請求項１乃至７の何れか１項に記載の情報処理装置。
【請求項９】
前記統合処理部は、前記中間層出力に対して正規化処理を行う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】
前記一部は、あらかじめ決められた複数の候補の中から手動または自動で選択されることを特徴とする請求項４乃至６の何れか１項に記載の情報処理装置。
【請求項１１】
前記一部は、前記認識対象の一部を検出する検出器の検出結果に基づいて動的に決定されることを特徴とする請求項４乃至６の何れか１項に記載の情報処理装置。
【請求項１２】
前記注目領域は、少なくとも人体の部位の位置を検出する検出器の検出結果に基づいて決定される領域を含むことを特徴とする請求項１乃至１１の何れか１項に記載の情報処理装置。
【請求項１３】

前記注目領域は、少なくとも物体の位置を検出する検出器の検出結果に基づいて決定される領域を含むことを特徴とする請求項１乃至１１の何れか１項に記載の情報処理装置。
【請求項１４】
学習対象の情報を含む学習データを入力として畳み込みニューラルネットワークの認識処理の結果を取得する認識処理部と、
前記学習データおよび／または前記畳み込みニューラルネットワークの中間層出力に対する注目領域を設定する注目領域設定部と、
前記注目領域中の前記学習データおよび／または前記中間層出力に対して前記認識処理よりも詳細な詳細認識処理を行う詳細認識部と、
前記詳細認識処理の結果と、前記中間層出力と、を統合処理する統合処理部と、
前記統合処理の結果を前記中間層出力として前記畳み込みニューラルネットワークに入力する中間入力処理部と、
前記認識処理部、前記注目領域設定部、前記詳細認識部、前記統合処理部、前記中間入力処理部、のいずれか一つ以上に関する学習処理を行う学習部と
を備えることを特徴とする情報処理装置。
【請求項１５】
前記詳細認識部が有する認識器が前記畳み込みニューラルネットワークであることを特徴とする請求項１４に記載の情報処理装置。
【請求項１６】
前記詳細認識部は少なくとも二つの認識器を備えることを特徴とする請求項１４又は１５に記載の情報処理装置。
【請求項１７】
前記注目領域設定部は、前記学習対象の一部に基づいて前記注目領域を設定し、
前記詳細認識部は、前記一部の種類ごとに認識器を備えることを特徴とする請求項１４乃至１６の何れか１項に記載の情報処理装置。
【請求項１８】
前記一部は少なくとも人体の部位に対応する情報を含むことを特徴とする請求項１７に記載の情報処理装置。
【請求項１９】
前記学習データは画像を含み、
前記統合処理部は、複数の前記一部の前記画像における位置に基づいて前記統合処理を行うことを特徴とする請求項１７又は１８に記載の情報処理装置。
【請求項２０】
前記認識処理部は、前記学習データに含まれる物体または現象に関する異常な状態を検知することを特徴とする請求項１４乃至１９の何れか１項に記載の情報処理装置。
【請求項２１】
前記認識処理部および前記詳細認識部は、マルチタスク学習に基づいて前記認識処理を行うことを特徴とする請求項１４乃至２０の何れか１項に記載の情報処理装置。
【請求項２２】
前記統合処理部は、前記中間層出力に対して正規化処理を行うことを特徴とする請求項１４乃至２１の何れか１項に記載の情報処理装置。
【請求項２３】
前記一部は、あらかじめ決められた複数の候補の中から手動または自動で選択されることを特徴とする請求項１７乃至１９の何れか１項に記載の情報処理装置。
【請求項２４】
前記一部は、前記学習対象の一部を検出する検出器の検出結果に基づいて動的に決定されることを特徴とする請求項１７乃至１９の何れか１項に記載の情報処理装置。
【請求項２５】
前記注目領域は、少なくとも人体の部位の位置を検出する検出器の検出結果に基づいて決定される領域を含むことを特徴とする請求項１４乃至２４の何れか１項に記載の情報処

理装置。
【請求項２６】
前記注目領域は、少なくとも物体の位置を検出する検出器の検出結果に基づいて決定される領域を含むことを特徴とする請求項１４乃至２４の何れか１項に記載の情報処理装置。
【請求項２７】
前記統合処理は、少なくとも加法演算を含むことを特徴とする請求項１乃至２６の何れか１項に記載の情報処理装置。
【請求項２８】
前記加法演算は、注目領域に基づいて部分てきに加法演算が適用されることを特徴とする請求項２７記載の情報処理装置。
【請求項２９】
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認識処理部が、認識対象の情報を含む認識対象データを入力として畳み込みニューラルネットワークの認識処理の結果を取得する工程と、
前記情報処理装置の注目領域設定部が、前記認識対象データおよび／または前記畳み込みニューラルネットワークの中間層出力に対する注目領域を設定する工程と、
前記情報処理装置の詳細認識部が、前記注目領域中の前記認識対象データおよび／または前記中間層出力に対して前記認識処理よりも詳細な詳細認識処理を行う工程と、
前記情報処理装置の統合処理部が、前記詳細認識処理の結果と、前記中間層出力と、を統合処理する工程と、
前記情報処理装置の中間入力処理部が、前記統合処理の結果を前記中間層出力として前記畳み込みニューラルネットワークに入力する工程と、
前記情報処理装置の出力部が、前記認識処理の結果を出力する工程と
を有することを特徴とする情報処理方法。
【請求項３０】
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認識処理部が、学習対象の情報を含む学習データを入力として畳み込みニューラルネットワークの認識処理の結果を取得する工程と、
前記情報処理装置の注目領域設定部が、前記学習データおよび／または前記畳み込みニューラルネットワークの中間層出力に対する注目領域を設定する工程と、
前記情報処理装置の詳細認識部が、前記注目領域中の前記学習データおよび／または前記中間層出力に対して前記認識処理よりも詳細な詳細認識処理を行う工程と、
前記情報処理装置の統合処理部が、前記詳細認識処理の結果と、前記中間層出力と、を統合処理する工程と、
前記情報処理装置の中間入力処理部が、前記統合処理の結果を前記中間層出力として前記畳み込みニューラルネットワークに入力する工程と、
前記情報処理装置の学習部が、前記認識処理、前記注目領域の設定、前記詳細認識処理、前記統合処理、前記入力、のいずれか一つ以上に関する学習処理を行う工程と
を備えることを特徴とする情報処理方法。
【請求項３１】
コンピュータを、請求項１乃至２８の何れか１項に記載の情報処理装置の各部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、認識対象を認識するための技術に関するものである。
【背景技術】
【０００２】
映像等のデータから、例えば物体およびその状態を認識するために、学習データを用い

て認識器を学習し、該認識器を用いて該認識を行う装置や方法が知られている。このような方法の一つとして、Ｃｏｎｖｏｌｕｔｉｏｎａｌ Ｎｅｕｒａｌ Ｎｅｔｗｏｒｋ（ＣＮＮ）がある。

【０００３】

ＣＮＮは、物体認識・行動認識・シーン認識など、様々な応用を目的として近年用いられている。行動認識については、例えば非特許文献１では、それぞれＲＧＢ画像とオプティカルフローとを入力する二つのストリームからなるＣＮＮを用いて行動認識を行うアーキテクチャが提案されている。また詳細な行動を認識するために、非特許文献２では、人体パーツの局所領域ごとにＣＮＮを用いて特徴量を抽出する方法に関する技術が提案されている。

【先行技術文献】

【非特許文献】

【０００４】

【非特許文献１】Two-Stream Convolutional Networks for Action Recognition in Videos. K. Simonyan, A. Zisserman. NIPS, 2014

【非特許文献２】P-CNN: Pose-based CNN Features for Action Recognition. Guilhem Cheron, Ivan Laptev, and Cordelia Schmid. ICCV, 2015

【発明の概要】

【発明が解決しようとする課題】

【０００５】

入力データの詳細な認識をより高精度に行うために、認識能力がさらに向上した認識システムが望まれる。本発明はこのような点に鑑みてなされたものであり、認識対象の認識精度を向上させるための技術を提供する。

【課題を解決するための手段】

【０００６】

本発明の一様態は、認識対象の情報を含む認識対象データを入力として畳み込みニューラルネットワークの認識処理の結果を取得する認識処理部と、前記認識対象データおよび／または前記畳み込みニューラルネットワークの中間層出力に対する注目領域を設定する注目領域設定部と、前記注目領域中の前記認識対象データおよび／または前記中間層出力に対して前記認識処理よりも詳細な詳細認識処理を行う詳細認識部と、前記詳細認識処理の結果と、前記中間層出力と、を統合処理する統合処理部と、前記統合処理の結果を前記中間層出力として前記畳み込みニューラルネットワークに入力する中間入力処理部と、前記認識処理の結果を出力する出力部とを有することを特徴とする。

【発明の効果】

【０００７】

本発明の構成により、認識対象の認識精度を向上させることができる。

【図面の簡単な説明】

【０００８】

【図１】異常検知システムの構成例を示すブロック図である。

【図２】異常検知システムの構成例を示すブロック図である。

【図３】異常検知システムの構成例を示すブロック図である。

【図４】異常検知システムの構成例を示すブロック図である。

【図５】異常検知システムが行う処理のフローチャート。

【図６】ステップＳ５０１における処理の詳細を示すフローチャート。

【図７】異常検知システムが行う処理のフローチャート。

【図８】前段処理部１１の構成例を示すブロック図。

【図９】前段処理部１１が行う処理のフローチャート。

【図１０】１フレームの画像の一例を示す図。

【図１１】認識部１２の構成例を示すブロック図。

【図１２】認識部１２が行う処理のフローチャート。

【図１３】ＣＮＮの一例を示す図。
【図１４】注目領域制御部２１の構成例を示すブロック図。
【図１５】認識部２２の構成例を示すブロック図。
【図１６】実行部２２２の構成例を示すブロック図。
【図１７】出力部２３の構成例を示すブロック図。
【図１８】学習時の動作のフローチャート。
【図１９】第１のＣＮＮ及び第２のＣＮＮの構成とデータフローに関する模式図。
【図２０】パーツのデータの形成例を示す図。
【図２１】第１のＣＮＮ及び第２のＣＮＮの構成とデータフローに関する模式図。
【図２２】認識部２２ａの構成例を示すブロック図。
【図２３】コンピュータ装置のハードウェア構成例を示すブロック図。
【図２４】第１のＣＮＮ及び第２のＣＮＮの構成とデータフローに関する模式図。
【図２５】第１のＣＮＮ及び第２のＣＮＮの構成とデータフローに関する模式図。
【図２６】第１のＣＮＮ及び第２のＣＮＮ及び後段のモデルの構成とデータフローに関する模式図。
【図２７】第１のＣＮＮ及び第２のＣＮＮのデータフローに基づいて得られる中間処理結果の可視化例。
【図２８】第１のＣＮＮ及び第２のＣＮＮの構成とデータフローに関する模式図。
【発明を実施するための形態】
【０００９】
以下、添付図面を参照し、本発明の実施形態について説明する。なお、以下説明する実施形態は、本発明を具体的に実施した場合の一例を示すもので、特許請求の範囲に記載した構成の具体的な実施例の１つである。
【００１０】
［第１の実施形態］
本実施形態では、認識器としてのＣｏｎｖｏｌｕｔｉｏｎａｌ Ｎｅｕｒａｌ Ｎｅｔｗｏｒｋ（ＣＮＮ）を学習する例を示す。なお、具体例は後述するが、本実施形態では、映像中の対象人物の行動に関する認識を行うＣＮＮと、さらに詳細な認識を行うＣＮＮ（詳細ＣＮＮ）と、を用いる場合の構成及び動作の例を示す。
【００１１】
本実施形態では認識処理を行う際に、通常の特徴量だけでなく、認識に寄与しうる注目領域の詳細な特徴量（詳細特徴量）をも抽出し、該抽出した詳細特徴量を通常の特徴量に順次統合しながら認識処理を行う。具体的には、通常の特徴を認識する第１のＣＮＮの中間層特徴マップに、詳細な特徴を認識する第２のＣＮＮの特徴量（詳細特徴量）を統合する。このとき、第１のＣＮＮの中間層特徴マップ上の注目領域に対応する領域に対して詳細特徴量を統合することで、詳細特徴量の統合を実現することができる（ROI Embedding）。ここで「統合」とは、一つ以上の詳細特徴量を第１のＣＮＮの中間層特徴マップに統合することを指し、例えば足し合わせや、代入、非線形変換等を通して、中間層特徴マップ上に詳細特徴量を反映させることを指す。「ROI Embedding」は、時々刻々と変化する注目領域の位置・大きさ・形等に対応して、第１のＣＮＮの特徴量と詳細特徴量との統合処理のためのデータフローを変更し、全体のニューラルネットワーク（ＮＮ）の構造を変化させる、という特性がある。特に、本実施形態では、注目領域を人体の部位（パーツ）として、各パーツの詳細な特徴量を抽出し、全身の特徴を認識する第１のＣＮＮの中間層特徴マップ上に順次統合処理する例を示す。
【００１２】
本実施形態では、監視カメラなどの撮像装置により撮影された映像中の物体や現象に係る異常を検知するためのシステムである異常検知システムを例にとり説明する。本実施形態に係る異常検知システムでは、監視対象を撮像装置で撮影し、該撮影により得られる映像に基づいて、監視対象に異常があるか否かを判定する。そして異常検知システムは、監視対象に異常があると判定した場合は、警備室等の監視センタに常駐する監視者に警告す

るための処理を行う。この監視対象には、例えば、一般家庭の屋内及び屋外、又は病院、駅などの公共施設が含まれる。

【００１３】

本実施形態に係る異常検知システムの動作は、「学習時:Forward Stage」、「学習時:Backward Stage」、「学習時:Final Stage」、「検出時」、の４つのステージからなる。これら４つのステージのうち、最初の３ステージは「学習時」の動作を示しており、４つ目のステージは学習結果に基づいて上記の異常検知処理を行う際（「検出時」）の動作を示している。

【００１４】

図１は、本実施形態に係る異常検知システムが有する構成のうち、ＣＮＮを学習させる動作ステージである「学習時:Forward Stage」に関連する構成の一例を示すブロック図である。認識装置１０と詳細認識装置２０との間は電子回路を介して接続されても良いし、記憶装置などの装置を介して接続されても良いし、ネットワーク（無線ネットワーク及び有線ネットワークのうち１つ以上により構成される）を介して接続されても良い。このネットワークには、例えば、携帯電話回線網やインターネットが適用できる。なお、以下に説明する装置や機能部についても同様に、装置間や機能部間は如何なる接続形態で接続されても良い。

【００１５】

図２は、本実施形態に係る異常検知システムが有する構成のうち、ＣＮＮを学習させる動作ステージである「学習時:Backward Stage」に関連する構成の一例を示すブロック図である。また、図３は、本実施形態に係る異常検知システムが有する構成のうち、ＣＮＮを学習させる動作ステージである「学習時:Final Stage」に関連する構成の一例を示すブロック図である。図３に示した構成では、図１，２に示した構成に異常判定装置３０が加わっている。

【００１６】

異常検知システムのＣＮＮの学習時における動作について、図５のフローチャートに従って説明する。ステップＳ５０１では、認識装置１０は、学習データを入力データとして第１のＣＮＮの学習を行い、詳細認識装置２０は、学習データと、認識装置１０からの出力である「第１のＣＮＮの出力」と、を入力データとして、第２のＣＮＮの学習を行う。ここで、第１のＣＮＮは、認識装置１０が使用するＣＮＮである。また、第２のＣＮＮは、詳細認識装置２０が使用するＣＮＮであり、第１のＣＮＮの出力等を受け取って詳細な認識処理を行うＣＮＮで、詳しくは後述する。また、第１のＣＮＮは、第２のＣＮＮの（中間の）詳細認識処理結果をさらに受け取って認識処理を行うように動作するものとし、詳しくは後述する。なお、ステップＳ５０１の動作は、ステージ「学習時:Forward Stage」における動作と、ステージ「学習時:Backward Stage」における動作と、を含んでいる。各ステージでの詳細な動作については後述する。

【００１７】

ステップＳ５０２では、異常判定装置３０は、認識装置１０から認識結果を受け取り、異常検知に係る閾値（以下、異常閾値と称する）を求める。ステップＳ５０２の動作は、ステージ「学習時:Final Stage」の動作に対応し、詳細な内容については後述する。

【００１８】

ステップＳ５０３では、認識装置１０による学習処理で得られた第１のＣＮＮのモデル、詳細認識装置２０による学習処理で得られた第２のＣＮＮのモデル、異常判定装置３０が求めた異常閾値、のそれぞれの保存処理が行われる。第１のＣＮＮのモデル及び第２のＣＮＮのモデルはそれぞれ、認識装置１０及び詳細認識装置２０によって、記憶部Ｍ１に保存され、異常閾値は、異常判定装置３０によって記憶部Ｍ２に保存される。ここで、ＣＮＮ（第１のＣＮＮ及び第２のＣＮＮのそれぞれ）のモデルとは、ＣＮＮを規定するパラメータを保持するデータセットである。ＣＮＮのモデルは、ＣＮＮにおけるニューロン間の結合荷重を規定するデータ、ＣＮＮのネットワーク構造を規定するデータ、ニューロン応答関数を規定するデータ、などを含む。

【００１９】
図４は、本実施形態に係る異常検知システムが有する構成のうち、異常検出時の動作に係る動作ステージである「検出時」に関連する構成の一例を示すブロック図である。図４に示した構成では、図３に示した構成に、ＰＣ（Ｐｅｒｓｏｎａｌ　Ｃｏｍｐｕｔｅｒ）のディスプレイやタブレットＰＣ、スマートフォン、フューチャーフォン等が適用可能な端末装置４０が加わっている。
【００２０】
異常検知システムのステージ「検出時」における動作について、図７のフローチャートに従って説明する。ステップＳ７０１では、認識装置１０及び詳細認識装置２０はそれぞれ、第１のＣＮＮを用いた検出対象データの認識処理（第１のＣＮＮの認識処理）、第２のＣＮＮを用いた検出対象データの認識処理（第２のＣＮＮの認識処理）を行う。このとき用いられる第１のＣＮＮ及び第２のＣＮＮは、上記の学習時に得られたモデルを用いるものとするが、異常検知システムの学習時に得られたモデルを用いず、その他のモデルを用いてもよい。
【００２１】
ステップＳ７０２では、異常判定装置３０は、検出対象データに対する第１のＣＮＮの認識処理の結果に対して異常判定を行う。このとき、異常判定を行う際の閾値として、学習時に得られた異常閾値を用いるものとするが、異常検知システムの学習時に得られた閾値を用いず、その他の閾値（例えば、ユーザが指定した閾値など）を用いてもよい。このとき、どのように閾値判定を行うかに関しては、後述する。
【００２２】
ステップＳ７０３では、端末装置４０は、異常判定装置３０による異常判定の結果を受け取り、該受け取った判定結果に基づく表示処理を行う。この表示処理の詳細については後述する。
【００２３】
次に、各装置の動作に関して説明を行う。ここで示すのは、各装置の動作の大まかな内容と処理の流れの順番である。それぞれの装置を構成する各機能部の詳細な構成・動作に関しては、機能部の説明をする際にあらためて述べるものとする。
【００２４】
まず、学習時の動作に関する説明を行う。学習時の動作の流れはすでに図５を用いて述べたとおりであるから、ここでは、図５の説明を補完する形で図６に基づいて説明を行う。
【００２５】
ステップＳ５０１では上記の通り、ステージ「学習時:Forward Stage」及びステージ「学習時:Backward Stage」の学習処理が実行される。ステップＳ５０１の処理の詳細について、図６のフローチャートに従って説明する。
【００２６】
ステップＳ６０１では、図１の構成により、ステージ「学習時:Forward Stage」の学習処理が実行される。まず前段処理部１１は、記憶部Ｄ１から学習データを受け取る。本実施形態では、記憶部Ｄ１には学習データとして、監視カメラなどの撮像装置が撮影した人体を含む映像が保存されているものとするので、前段処理部１１は、この映像を学習データとして記憶部Ｄ１から取得する。そして前段処理部１１は、映像中の人体の領域（人体領域）を抽出する等の前処理を行ったのち、学習データを第１のＣＮＮ用の学習データとして整形して認識部１２に送る（詳細は後述する）。認識部１２は、前段処理部１１から受け取った学習データに基づいて、第１のＣＮＮの認識処理を行う。第１のＣＮＮの認識処理とは、第１のＣＮＮによってデータを認識するための一連のプロセスのことを指す。なお、本実施形態に係る認識処理は、第１のＣＮＮの認識処理と第２のＣＮＮの認識処理とが相互に依存する構成になっており、具体例を以降で説明する。認識部１２は、第１のＣＮＮの認識処理において得られる該第１のＣＮＮにおける規定の中間層の出力（中間認識情報）を、注目領域制御部２１に送る。また、前段処理部１１は、認識部１２が受け取

った学習データ（後述する教師データを含む）を、注目領域制御部２１に送る。注目領域制御部２１は、受け取った学習データ及び第１のＣＮＮにおける規定の中間層の出力から、第２のＣＮＮが認識する対象を設定し、該設定した対象を規定する詳細認識対象データ（詳細認識対象情報）を認識部２２に送る。詳細認識対象データの具体的な作成方法については後述する。本実施形態では、詳細認識対象データとは、認識対象となる人物の行動認識を行うために着目すべき人体のパーツに関するデータを保持するものとする。注目領域制御部２１は、認識対象となる人物のパーツの領域（注目領域）を規定する領域データ（注目領域情報）を、出力部２３及び認識部１２に送る。認識部２２は、受け取った詳細認識対象データに対して認識処理を行い、該認識処理の結果（詳細認識情報）を詳細認識処理結果として出力部２３に送る。出力部２３は、認識部２２による認識の結果を認識部１２に出力するために、上記の受け取った領域データに基づいて、詳認識の結果を認識部１２に送る。すなわち、第１のＣＮＮの規定の中間層の出力の領域に対して送り返す形で詳細認識情報を送信することになる（この点に関してはバリエーションがあり、詳細は後述する）。認識部１２は、第１のＣＮＮの中間層出力と、出力部２３から受け取った詳細認識情報と、に基づいて、第１のＣＮＮの認識処理を続行する。

【００２７】

その際、さらに第２のＣＮＮを利用して詳細な認識処理を行ってもよい。なお認識部１２は、受け取った領域データを、以降のステージ「学習時:Backward Stage」で利用する。最後に認識部１２は、あらかじめ定めた損失関数に基づいて、処理を行う。例えば、学習データに含まれる教師データに基づいて、二乗誤差によって誤差を算出する方法などがありえる。なお、二乗誤差ではなく、その他の公知な損失関数を用いてもよい。

【００２８】

次に、ステップＳ６０２では、図２の構成により、ステージ「学習時:Backward Stage」の学習処理が実行される。ここでは、認識部１２による認識処理の結果に対する評価結果をもとにモデルの更新を行う。学習アルゴリズムには、例えば誤差逆伝播法を用いることができる。誤差逆伝播法は、評価結果の誤差をニューラルネットワークの上層から下層に伝播してゆく形で学習を行う手法であり、詳細は後述する。まず、認識部１２は、誤差情報をもとにモデルの更新を行う。また、第２のＣＮＮの出力が入力された第１のＣＮＮの領域に関する誤差情報を、第２のＣＮＮに対して伝播する。このとき、認識部１２は、領域データに基づいて第１のＣＮＮの注目領域を特定し、出力部２３に該注目領域の誤差情報を送るものとする。なお、ここで述べる誤差とは、誤差逆伝播法による学習の過程において、第１のＣＮＮの上層ネットワークから注目領域に逆伝播された誤差を示す。さらに出力部２３は、認識部２２に誤差情報を送る。このとき、第２のＣＮＮのどの部分に誤差情報を送るかどうかを決定する必要がある。ここでは、注目領域に対して詳細情報を出力した第２のＣＮＮの出力ニューロンに対して誤差情報を送るものとする。すなわち、ステージ「学習時:Forward Stage」の際に入力データが認識処理された経路の逆順をたどるように、誤差情報が伝播するものとする。認識部２２は、受け取った誤差情報に基づいて第２のＣＮＮのモデルを更新する（詳細は後述する）。また、受け取った誤差情報を、さらに認識部１２に対して逆伝播し、認識部１２は、受け取った誤差情報に基づいて第１のＣＮＮのモデルを更新する。以上のように、誤差逆伝播法に基づいて第１のＣＮＮ及び第２のＣＮＮを更新することができる。

【００２９】

次に、ステップＳ６０３では、学習が完了したか否かの判定が行われる。この判定の基準については後述する。学習が完了していないと判定された場合は、ステップＳ６０１に戻り、学習が完了したと判定された場合は、図６の動作は終了し、処理はステップＳ５０２に移行する。

【００３０】

図５に戻って、次にステップＳ５０２では、ステージ「学習時:Final Stage」として、ステップＳ６０１及びステップＳ６０２において学習した第１のＣＮＮ及び第２のＣＮＮの認識結果に基づいて、異常検知をするための閾値が決定される。認識部１２は、学習デ

ータに対する認識結果を異常判定部３１に送る。異常判定部３１は、認識部１２の認識結果に基づいて、認識対象が異常かどうかを判定するための基準である閾値を決定する。このときの閾値の決定方法には、公知の既存の手法を用いればよい。例えば、第１のＣＮＮが認識対象の行動を認識するように学習が行われており、認識対象が「転倒」したか否かを認識することができるとする。このとき、「転倒」クラスが異常と判定すべきクラスであるものとして、認識対象の行動が「転倒」クラスであることを示す確率が５０％以上であったら、異常であると判定するとしてもよい。このとき「５０％」が異常閾値である。このようにあらかじめ固定の閾値を与える方法以外のアプローチとして、閾値を評価用データによって決定してもよい。具体的には、学習データセットとは異なる評価用データセット（例えば映像）を用意する。このとき、学習済みの第１のＣＮＮを用いて各評価データに対する「転倒」の確率を求めることで、例えば真に「転倒」であるすべての評価データを「転倒」であると判定できる閾値を決定することができ、このような閾値を異常閾値として設定してもよい。その他の方法として、第１のＣＮＮの特徴量に基づいて正常の範囲のモデル（正常モデル）をさらに学習し、学習されたモデル（正常モデル）に基づいて正常か否かを判定してもよい。このとき用いる第１のＣＮＮの特徴量は、例えば最終層より一層前の層の出力（中間層出力）を特徴量として用いてもよい。また、このとき用いる正常モデルを学習する方法には如何なる方法を採用しても良く、例えば特開2014-203289号公報に記載の方法を用いることができる。ここで決定した閾値や、正常モデルなど、異常の判定に用いる特徴量は、記憶部Ｍ２に格納される。

【００３１】

次に、異常検知システムのステージ「検出時」における動作について、図７のフローチャートに従って説明する。ステップＳ７０１では、第１のＣＮＮの認識処理、第２のＣＮＮの認識処理が行われる。先ず、前段処理部１１は、記憶部Ｄ２から検出対象データとして、監視カメラなどの撮像装置により撮影された人体を含む映像を取得する。そして前段処理部１１は、学習時と同様に映像から人体領域を抽出するなどの前処理を行ったのち、検出対象データを第１のＣＮＮ用の検出対象データとして整形（詳細は後述する）して、認識部１２に送る。そして認識部１２及び認識部２２はそれぞれ、この検出対象データに基づいて第１のＣＮＮの認識処理及び第２のＣＮＮの認識処理を行う。この処理の詳細は、上記のステージ「学習時:Forward Stage」と同様であるため、その説明は省略する。

【００３２】

次にステップＳ７０２では、異常判定部３１は、検出対象データに対する認識部１２の認識結果に基づいて、異常かどうかの判定を行う。このとき異常判定部３１は、上記のステージ「学習時:Final Stage」で記憶部Ｍ２に格納した異常閾値（あるいは正常モデル）を受け取り、異常の判定に用いるものとする。

【００３３】

次にステップＳ７０３では、端末装置４０は、異常判定部３１による異常判定結果を受け取り、受け取った結果に基づく表示を表示部４１に表示させるための表示処理を行う。表示部４１は、検出対象が異常であることを示す異常判定結果が送られてきた場合、警告を示す表示を行ってもよい。例えば、監視カメラなどの撮像装置による撮影映像上で異常と判定された検出対象の箇所を強調表示してもよい。このとき、異常な箇所を強調表示するには、映像上の異常な箇所を特定する必要がある。そのためには、例えば入力される検出対象データに画面座標データを付与しておき、異常判定結果に応じてその座標データを利用してもよい。

【００３４】

なお、異常検知システムは、検出対象が異常である場合における警告動作として警告を示す表示を行うことに加えて若しくは代えて、他の処理を行うようにしても良い。例えば、異常検知システムにランプとサイレンが備わっている場合はサイレンにより警告音を鳴らすと共にランプを点滅させても良い。

【００３５】

次に、次の映像が残っている場合には、処理はステップＳ７０４を介してステップＳ７

０１に戻る。一方、次の映像が残っていなければ、図７のフローチャートに従った処理は終了する。

【００３６】

次に、上記の各装置が有する各機能部の構成及び動作について説明する。図８は、上記の前段処理部１１が有する機能部のうち、学習時及び検出時に関連する構成の一例を示すブロック図である。また、図１１は、上記の認識部１２が有する構成のうち、学習時に関連する構成の一例を示すブロック図である。図１４は、上記の注目領域制御部２１が有する構成のうち、学習時に関連する構成の一例を示すブロック図である。図１５は、上記の認識部２２が有する構成のうち、学習時に関連する構成の一例を示すブロック図である。図１６は、図１５に示す認識部１２の実行部２２２が有する構成において、学習時に関連する構成の一例を示すブロック図である。図１７は、上記の出力部２３が有する構成のうち、学習時に関連する構成の一例を示すブロック図である。

【００３７】

次に、前段処理部１１（図８）の学習時の動作について、図９のフローチャートに従って説明する。ステップＳ９０１では、抽出部１１１は、記憶部Ｄ１から学習データを読み込む。本実施形態では、この学習データは、監視カメラなどの撮像装置により撮影された人体を含む映像を含むとする。抽出部１１１は、映像から人体領域を抽出する。映像が動画像である場合には、各フレームの画像から人体領域を抽出することになる。ここで、動画像としての映像（監視映像）における、ある１フレームの画像の一例を図１０に示す。図１０の画像１００１は、ある交差点における監視映像における１フレーム分の画像の一例であり、１００２～１００５は、該画像１００１に含まれているオブジェクトである。１００６，１０１０，１００９，１００７はそれぞれ、オブジェクト１００２～１００５について抽出部１１１が抽出したＢｏｕｎｄｉｎｇ　Ｂｏｘであり、これらのＢｏｕｎｄｉｎｇ　Ｂｏｘに囲われた部分画像それぞれが人体領域に対応している。なお、ここで示したＢｏｕｎｄｉｎｇ　Ｂｏｘはあくまでも人体領域が抽出された際の具体例の一つであり、例えば後述する背景差分法によって撮像されたオブジェクトの輪郭に沿った小領域を抽出してもよい。１０９９は信号を示している。

【００３８】

映像からこのような人体領域を抽出するための方法は複数存在し、如何なる方法を採用しても良い。このような方法には、例えば背景差分法、物体検出・追尾法、領域分割法の三つがある。

【００３９】

抽出部１１１は、人体のように監視対象のオブジェクトがあらかじめ既知である場合は、ターゲットのオブジェクトのみを検出・追尾する目的に絞られた物体検出・追尾法が比較的適していると考えられる。物体検出・追尾法には、例えば以下の文献１に記載の方法がある。

【００４０】

文献１　…　Real-Time Tracking via On-line Boosting, H. Grabner, M. Grabner and H. Bischof, Proceedings of the British Machine Conference, pages 6.1-6.10. BMVA Press, September 2006

さらに抽出部１１１は、学習データ（監視映像）にあらかじめ付与された教師データを利用して、人体領域に対して教師データを付与する。例えば、矩形の人体領域を抽出する場合は人体領域のBounding Boxを定義し、該Bounding Boxに教師データを付与することができる。また、例えば人体の輪郭に沿って人体領域を抽出する場合は、画像に対する人体領域のマスクを定義し、該マスクに教師データを付与することができる。教師データは対象をどのように分類すべきかを示すラベルである。ラベルの種類や、どのような対象にどのようなラベルを付与するか、ということは問題に依存するため、異常検知システムを使用または導入するユーザがあらかじめ決定し、学習データに対して教師データを付与しておくものとする。ここで、学習データにあらかじめ付与する教師データは、例えば撮像された被写体の画像上の領域に対して、ユーザが付与することができる。より具体的には、

例えば歩行者であるオブジェクト１００３の領域をユーザが手動で指定し、その領域に対して例えば歩行者を示すラベルを付与することができる。このとき、抽出部１１１は、抽出された人体領域と、付与された教師データの領域と、が重畳している場合、その人体領域に対して最も大きな面積の割合で重畳された教師データを付与してよい。なお、必ず上記のやり方で教師データを付与しなければならないわけではなく、例えばあらかじめ人体領域を抽出しておき、それぞれの人体領域に対して、ユーザが教師データを付与してもよい。なお、本実施形態では人体領域を抽出する例を示したが、抽出する領域は人体領域に限らず、他の対象物の領域を抽出するようにしても良い。また、領域を抽出せずに画像全体を認識処理の対象としてもよい。どのような領域を抽出し、認識処理の対象とするかは問題依存であるため、問題に応じて設定する必要がある。最後に、抽出部１１１は、抽出した人体領域（画像）を前処理部１１３と検出部１１２とに送る。なお、説明を簡略化するために、ステップＳ９０１では、異常検知システムで用いられるすべての学習データを一度に受け取り、受け取った学習データから人体領域を抽出し、抽出されたすべての人体領域を以降の処理の対象とする。なお、学習データのデータサイズが非常に大きい場合は、すべてのデータを通信により一度に取得することは難しいため、その場合は一部ずつ学習データを取得するようにしてもよい。以降の説明では、特に注意書きを記載する場合（後述するMinibatch等に関して記述している場合等）を除いては、すべての学習データに対して一度に処理・通信を行うものとする。

【００４１】

ステップＳ９０２では、検出部１１２は、抽出部１１１から人体領域を受け取る。ここで受け取る人体領域の例を、図１０下部に１０１１として示す。人体領域１０１１における１００３は、図１０上部の画像１００１中のオブジェクト（人物）１００３と同一の人物であることを示している。１０１２は頭部領域、１０１３は右手領域、１０１４は右足領域、１０１５は左手領域、１０１６は左足領域であり、それぞれのパーツ領域は、検出部１１２が検知したパーツ領域の一例である。ここで検出する対象のパーツ領域は、例えば認識すべき問題に応じてあらかじめ決めてもよい。例えば、現象の１つである万引き行為を検知するために、手を検出対象のパーツとしてもよい。ここでは、あらかじめ上記５つのパーツ領域を検出するものとする。パーツ領域を検出するための方法には如何なる方法を用いても良く、例えば以下の文献２に記載の方法がある。

【００４２】

文献２　…　Convolutional Pose Machines, Shih-En Wei, Varun Ramakrishna, Takeo Kanade and Yaser Sheikh, CVPR, 2016.

なお、ここでは、人体領域を抽出した後にパーツ領域の検出を行う例を示したが、これに限るものではなく、多数の人物のパーツ領域を特定したのち、人体領域を特定するようなプロセスを用いてもよい。

【００４３】

図９に戻って、次にステップＳ９０３では、検出部１１２は、ステップＳ９０２において検出部１１２が人体領域から検出したパーツ領域を抽出し、該抽出したパーツ領域を前処理部１１３に送る。

【００４４】

次にステップＳ９０４では、前処理部１１３は、認識処理の前処理を行う。本実施形態では、各人体領域（画像）を規定の画像サイズ（例えば２２４画素×２２４画素）に変形（縮小）したのち、平均画像を引く処理を行うものとする。ここで平均画像は、学習データにおける複数枚の画像のそれぞれの画素位置（ｘ、ｙ）の画素値の平均値を平均画像における画素位置（ｘ、ｙ）の画素値とすることで得られる画像である。また、前処理部１１３は、学習データ（学習画像）を水平方向に反転して水増しし、水増し前の画像と水増し後の画像とを、すべて学習画像として扱うこととする。

【００４５】

次にステップＳ９０５では、データ選択部１１４は、学習データの部分集合群（Minibatch集合）を作成する。Minibatchは、第１のＣＮＮ及び第２のＣＮＮのパラメータをＳｔ

ｏｃｈａｓｔｉｃ Ｇｒａｄｉｅｎｔ Ｄｅｓｃｅｎｔ（ＳＧＤ、確率的勾配降下法）によって繰り返し最適化する際に用いる学習データの集合である。詳細は後述するが、本実施形態では、ＣＮＮ等のパラメータを最適化する方法としてＳＧＤを用いる例を示すこととする。なお、Minibatchのサイズは例えば本実施形態では、同じ教師データのラベルを持つ５０個の学習画像をひとまとめにしたものとする。また、Minibatchは、学習データが重複なく属するように、かつすべての学習データが含まれるように作成するものとする。

【００４６】

次にステップＳ９０６では、データ選択部１１４は、Minibatch集合からランダムに一つMinibatchを選択し、認識部１２及び詳細認識装置２０に送る。このとき選択するMinibatchは、その学習のＩｔｅｒａｔｉｏｎの中でまだ使用していないMinibatchを選択するものとする。ここでＩｔｅｒａｔｉｏｎとは、一般的にはエポックとも言われ、Minibatch集合のすべてのMinibatchを何回学習したかを表すものとする。

【００４７】

次にステップＳ９０７では、データ選択部１１４は、学習処理の終了判定を行う。学習処理の終了判定の基準には様々な基準を適用することができ、例えば、Ｉｔｅｒａｔｉｏｎが５０以上になった場合には終了するとしてもよい。学習処理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Ｓ９０６に戻り、学習処理の終了条件が満たされている場合には、図９のフローチャートに従った処理は終了する。

【００４８】

次に、認識部１２（図１１）の学習時の動作について、図１２のフローチャートに従って説明する。ステップＳ１２０１では、初期化部１２１は、第１のＣＮＮのパラメータ（結合加重及びバイアス項）を初期化する。初期化にあたっては、第１のＣＮＮのネットワーク構造をあらかじめ決めておく必要がある。ここで用いるネットワーク構造や初期パラメータは特定のパラメータに限らず、例えば以下の文献３と同じものを用いてよいし、独自に定義したネットワーク構造を用いてもよい。ここで図１３に、本実施形態に適用可能なＣＮＮの一例を示す。

【００４９】

文献３ … A. Krizhevsky et al. "ImageNet Classification with Deep Convolutional Neural Networks", Advances in Neural Information Processing Systems 25 (NIPS), 2012.

図１３のニューラルネットワーク１３２０は、本実施形態に係る第１のＣＮＮのネットワーク構造の一例を示したものである。図１３に示す如く、ニューラルネットワーク１３２０は、入力層１３０１、ｃｏｎｖｏｌｕｔｉｏｎ１層１３０２、ｐｏｏｌｉｎｇ１層１３０３、ｃｏｎｖｏｌｕｔｉｏｎ２層１３０４、ｐｏｏｌｉｎｇ２層１３０５、Ｉｎｎｅｒｐｒｏｄｕｃｔ１層１３０６、Ｉｎｎｅｒｐｒｏｄｕｃｔ２層１３０７、出力層１３０８、を有している。また、２つの階層間の処理方法として、ｃｏｎｖｏｌｕｔｉｏｎ処理１３１０、ｐｏｏｌｉｎｇ処理１３１１、ＩｎｎｅｒＰｒｏｄｕｃｔ処理１３１２が設定されている。それぞれの処理の具体的な内容は周知であり、例えば上記の文献３と同様であるため、ここでは省略する。

【００５０】

ｃｏｎｖｏｌｕｔｉｏｎ処理１３１０では、畳み込みフィルタを用いてデータ処理を実行し、ｐｏｏｌｉｎｇ処理１３１１では、例えば、ｍａｘ ｐｏｏｌｉｎｇであれば局所的な最大値を出力する処理を行う。また、ＩｎｎｅｒＰｒｏｄｕｃｔ処理１３１２では、内積処理を実行する。

【００５１】

ここで、畳み込みフィルタは、学習後の各ニューロンの受容野に該当する結合加重（結合重み係数）の分布で定義される。また、図１３では、ｃｏｎｖｏｌｕｔｉｏｎ層およびｐｏｏｌｉｎｇ層には複数の特徴マップ（中間層特徴マップ）が存在し、入力層の画像上のピクセルに対応する位置には、複数のニューロンが存在する。例えば学習画像がＲＧＢ

形式である場合、ＲＧＢチャンネルに対応する３つのニューロンが存在する。また、撮像された映像の動き情報を持つＯｐｔｉｃａｌＦｌｏｗ画像であれば、画像の横軸方向と縦軸方向とをそれぞれ表現する２種類のニューロンが存在することになる。また、複数の画像を同時に入力として用いる場合は、入力画像の数に対応する分だけ入力層のニューロンを増やすことで対応することが可能である。本実施形態では、標準的なＲＧＢ画像を対象とする例を示すものとする。

【００５２】

なお、上記のステップＳ１２０１における初期化の方法にはバリエーションが存在する。初期化部１２１は、公知の手法で第１のＣＮＮのパラメータを初期化することができる。初期化のバリエーションには大きく分けて、データを用いて初期化する場合と、データを用いないで初期化する場合と、がある。ここでは、データを用いないで初期化する場合の簡単な方法として、平均０分散１の正規分布からランダムにサンプリングした値を用いて重みパラメータを初期化し、さらにバイアス項パラメータはすべて０で初期化するものとする。なお、データを用いて初期化する場合は、例えば画像認識などを用途としてあらかじめ学習（プレトレーニング）した第１のＣＮＮを、本実施形態に係る第１のＣＮＮの初期モデル（初期化された結合重みパラメータなど）として用いてもよい。そして初期化部１２１は、上記のようにして初期化した第１のＣＮＮのモデルを実行部１２２に送る。

【００５３】

次にステップＳ１２０２では、実行部１２２は、変数Ｉｔｅｒａｔｉｏｎの値を０に初期化する。次にステップＳ１２０３では、実行部１２２は、学習用のMinibatchを前段処理部１１から取得する。

【００５４】

次にステップＳ１２０４では実行部１２２は、第１のＣＮＮの認識処理を実行する。このとき、第１のＣＮＮの認識と第２のＣＮＮの認識とを連携して行うための処理を実行する（詳細は後述する）。

【００５５】

次にステップＳ１２０５ではＣＮＮ更新部１２３は、ステップＳ１２０４で得られた認識処理結果に基づいて、第１のＣＮＮのモデルを更新するための処理を行うために、認識誤差および誤差情報の伝播を行う。本実施形態では、第１のＣＮＮの学習を行うための方法として、誤差逆伝播法とＳＧＤとを組み合わせた方法を用いる。誤差逆伝播法とＳＧＤとを組み合わせた方法は周知であり、例えば、上記の文献３に記載の通り、Minibatchを選択し、第１のＣＮＮのパラメータを逐次更新するという手順を繰り返す方法である。

【００５６】

なお、誤差情報は認識処理のデータフローとは逆方向に伝播するのが一般的であり、ここでも第２のＣＮＮに対して誤差情報が伝播するものとする。ここで本実施形態においては、誤差逆伝播処理のためのデータフローは動的に変わりうることに注意が必要である。これについての詳細な説明は後述する。

【００５７】

次にステップＳ１２０６では、ＣＮＮ更新部１２３は、ステップＳ１２０５で伝播した誤差情報を用いて、第１のＣＮＮの更新を行う。次にステップＳ１２０７では、実行部１２２は、Minibatchをすべて学習に利用したか否かの判定を行う。すべてのMinibatchを学習に利用した場合は、ステップＳ１２０８に移行し、未だ学習に利用していないMinibatchが残っているのであれば、処理はステップＳ１２０３に移行する。

【００５８】

ステップＳ１２０８では、実行部１２２は、変数Ｉｔｅｒａｔｉｏｎの値に１を加算する。ステップＳ１２０９では、実行部１２２は、変数Ｉｔｅｒａｔｉｏｎの値（Ｉｔｅｒａｔｉｏｎ回数）が予め設定した上限値に達しているか否かを判定する。この判定の結果、Ｉｔｅｒａｔｉｏｎ回数が予め設定した上限値に達している場合は、第１のＣＮＮの学習を終了し、処理はステップＳ１２１０に移行する。一方、Ｉｔｅｒａｔｉｏｎ回数が予め設定した上限値に達していない場合は、処理はステップＳ１２０３に移行する。

【００５９】
なお多くの場合、ＮＮの学習停止条件は、「Ｉｔｅｒａｔｉｏｎ回数が予め設定した上限値に達した」であったり、学習曲線の勾配を用いて自動的に決めたり、そのどちらかを採用する。本実施形態では、学習の停止条件として、「Ｉｔｅｒａｔｉｏｎ回数が予め設定した上限値に達している」を採用するが、これに限るものではない。本実施形態では、「予め設定した上限値」を「２００００」とするが、この値に限らない。なお、Ｉｔｅｒａｔｉｏｎ回数が増えることによってＮＮの学習率を低下させる方法など、Ｉｔｅｒａｔｉｏｎ回数に基づく学習処理の工夫を導入してもよい。
【００６０】
次に、図１４～図１７の構成による学習時の動作について、図１８のフローチャートに従って説明する。ステップＳ１８０１では、初期化部２２１は第２のＣＮＮの初期化を行う。初期化の方法は公知の方法を用いることが可能であり、例えば上述したＣＮＮの初期化方法を用いてもよい。ここで初期化したモデルを操作部２２２１に送り、操作部２２２１は受け取ったモデルをストリーム２２２３が備える第２のＣＮＮのモデルとして読み込む。
【００６１】
ステップＳ１８０２では、受信部２１１は、データ選択部１１４から詳細認識対象データを受け取り、さらに受信部２１１は、実行部１２２から第１のＣＮＮの中間層出力を受け取る。受信部２１１は、受け取った詳細認識対象データを、実行部２２２の受信部２２２２に送る。受信部２２２２に送られた詳細認識対象データは、以降のステップで認識処理にかけられる。
【００６２】
ステップＳ１８０３では、設定部２１２は、前段処理部１１から注目領域情報を受け取る。ステップＳ１８０４では、受信部２２２２は、詳細認識対象データをストリーム２２２３に送る。ストリーム２２２３は、受け取った詳細認識対象データに対して、第２のＣＮＮを用いて認識処理を行う。ここで、具体例を示すために、図１９に第１のＣＮＮ及び第２のＣＮＮの構成とデータフローに関する模式図を例示する。図１９における例では、認識対象人物の行動を認識するために、第１のＣＮＮは認識対象人物の全身に関わる行動、第２のＣＮＮは認識対象人物の頭部に関する詳細な特徴をそれぞれ認識する役割があるとする。このとき第２のＣＮＮは、認識対象人物の頭部領域の画像と、第１のＣＮＮの認識した全身に関する中間層出力と、に基づいて、頭部に関わる動作を認識（特徴量を抽出）するものとする。ここで抽出された詳細な特徴量を、詳細認識処理結果として、第２のＣＮＮは第１のＣＮＮに送り、第１のＣＮＮは受け取った詳細認識結果をさらなる認識処理を行うための入力データとして用いる構成をとるとする。このとき第２のＣＮＮは、第１のＣＮＮの特徴マップ上の頭部領域の位置を考慮して、詳細認識結果を送ってもよい。このとき、第２のＣＮＮの出力する詳細認識結果は、頭部に関する詳細な特徴量である。この詳細な特徴量が、第１のＣＮＮの持つ認識対象人物の全身に関する中間層特徴マップの頭部領域に送られ、統合されることで、第１のＣＮＮの中間層特徴マップは頭部に関して詳細化されることが期待できる。このように詳細な特徴量を用いる構成によって、例えば人物の「キョロキョロ」を精度よく検知することが期待できる。１００３、１０１１、１０１２は、それぞれ図１０において示した認識対象人物の例、認識対象人物の全身領域の画像データ、認識対象人物の頭部領域の画像データ、を示している。１９０１は、頭部画像データと頭部領域の位置データを受け取る設定部２１２を表しており、この処理はステップＳ１８０３の処理を示している。１９０２は第２のＣＮＮの認識処理を行う対象の頭部領域画像データ、１９０３は第１のＣＮＮが認識処理を全身画像の例である。１９０４、１９１４はそれぞれストリーム２２２３が備える第２のＣＮＮの第一層、第二層の畳み込み処理を示している。１９０２は、頭部領域の画像データを第２のＣＮＮの第一層に入力するためにサイズの変形を行った画像を示しており、第２のＣＮＮの第一層に入力され、出力結果としての詳細認識処理結果１９０５が得られる。１９０５は第２のＣＮＮの第二層に入力され、以下同様に第２のＣＮＮの処理がなされる。ここまでが、ステップＳ

１８０４の処理である。なお、ここで説明した方法は一つの例であり、その他の方法を用いて第２のＣＮＮと第１のＣＮＮとを連携させてもよい。その他の例に関しては後述する。

【００６３】

ステップＳ１８０５では、ストリーム２２２３は、変形部２２２４に詳細認識処理結果を送る。変形部２２２４は、受け取った詳細認識処理結果を変形することで、第１のＣＮＮの中間層特徴マップに統合するための前段階としての処理を行う。このステップの処理の例は、図１９において、詳細認識処理結果１９０５が変形部２２２４である１９０８に送られ、変形結果１９０９が得られるという工程として表されている。このとき、どのような情報に基づいてどのように変形するかどうかはバリエーションがあり、後述する。

【００６４】

ステップＳ１８０６では、統合部２２２５は、変形部２２２４から変形された詳細認識処理結果を受け取り、これを入力データとして第１のＣＮＮの中間層特徴マップに統合するための処理を行う。この際に、どのように統合処理を行うかどうかのバリエーションが複数考えられるため、詳細は後述する。このステップの処理の例を、図１９において、変形結果１９０９と、第１のＣＮＮの認識処理結果（詳細認識対象データ）である１９０７を受け取り、詳細認識処理結果１９１２を出力する統合部２２２５（図１９では１９１０）として示す。

【００６５】

ステップＳ１８０７で、指定部２３１は、設定部２１２から注目領域情報を受け取る。出力部２３２は、指定部２３１が指定する第１のＣＮＮの注目領域に対して、詳細認識処理結果を送信する。ここで注目領域は、本実施形態においては認識対象人物の頭部領域である。第１のＣＮＮの特徴マップ（ある層のある特徴マップ）上の認識対象人物の頭部領域の大きさは、認識対象人物の原画像上での頭部の大きさに対して、第１のＣＮＮの処理等を経て変動しうる。例えば、認識対象人物の原画像上での頭部の大きさが既知であれば、第１のＣＮＮのある特定の層の特徴マップ上で表現される頭部の大きさは計算可能であり、設定部２１２はそのように頭部の注目領域の大きさを求めてもよい。ここで得られる頭部の注目領域の大きさの変動は、先述の変形処理によって吸収される（詳細認識処理結果の特徴マップ上の大きさが、その時々の注目領域の大きさにリサイズされる）。また、認識対象人物の頭部領域の位置は、人物の姿勢等によって変動しうる。この位置に関する変動は、設定部２１２が頭部領域の位置を指定することによって吸収される。以上のような変動は、時々刻々と認識対象人物の頭部領域が変化することで生じうる。上記の説明は、その変動を吸収するために、ニューラルネットワークの構造の一部が時々刻々と変化しうることを示している。例えば、頭部領域の位置が変わることによって詳細認識処理結果の送信先の領域が変動することは、ニューラルネットワークの結合先（ニューロン間の結合関係）が動的に変わることで実現できる。ニューラルネットワークの結合先（ニューロン間の結合関係）を動的に変える構成は、ソフトウェアで実現してもよいし、ハードウェアで実現してもよい。

【００６６】

例えば、以降で引用する文献６で用いられるVGG16というモデルでは、各層の特徴マップの大きさは、層のコンポーネントごとに定まっている（ここではプーリング層を挟むと層のコンポーネントが変わるという解釈のもと例示している）。具体的には、一段目から五段目までの層のコンポーネントにおいて、224x224、112x112、56x56、28x28、14x14というサイズの特徴マップとなる。また、入力画像サイズは224x224である。よって、各層のコンポーネントの特徴マップサイズと、入力画像のサイズとの比は、1、0.5、0.25、0.125、0.0625として定まる。例えばこの比に基づいて、ある特徴マップ上の注目領域の大きさを決定してもよい。

【００６７】

例えばソフトウェアで実現する場合、上述の通り、統合部２２２５（図１９では１９１０）が、詳細認識処理結果の送信先を認識時に変更してもよい。その場合、例えば認識時

において、同一のメモリ上に格納されたニューロン間でデータを送受信する際に、送受信先を決定するポインタを変更し、新たな送受信先となるニューロンを指定することが可能である。また、モデルの更新時には、認識時に用いたポインタを記憶しておき、誤差逆伝播法適用の際に誤差を伝播するデータフローを決定してもよい。その他の公知なソフトウェアによる方法によっても、ニューロン間の結合を変更してよい。

【００６８】

また、ハードウェアで実現する場合、例えばインターネットを介して通信機器間のデータ送受信先を切り替えることで、ニューロン間の結合を変更してもよい。例えばニューラルネットワークのモデルが複数の機器に分散して存在するとする。認識時において統合部２２２５（図１９では１９１０）が、送受信先となる機器のアドレスを変更することでニューロン間のデータの送受信先変更を実現することが可能である。その他の公知なハードウェアによる方法によっても、ニューロン間の結合を変更してもよい。

【００６９】

また、上述のハードウェア・ソフトウェアの実装を複合的に用いて、ニューロン間の結合を変更してもよい。また、ニューラルネットワーク中の結合重みを変更し、活性化状態を変化させることで、詳細認識処理結果の送信先を変更する処理を実現することができる。具体的には、詳細認識処理結果の送信先の候補に対して、ニューロン間の結合重みが存在するとする。送信先の候補とは、詳細認識処理結果を代入する第１のＣＮＮの特徴マップ中の領域の候補とも言える。この送信先の候補に対して、詳細認識処理結果が実際に送信されるかが事前には決定されておらず、したがって詳細認識処理結果の送信元と受信先となるニューロン間の結合重みは０、バイアス項も０の値を持つとする。そして、認識時に送信先の候補が選ばれ、送信先が確定するとする。このとき、確定した送受信間関係に基づいて、統合部２２２５（図１９では１９１０）が送受信の行われるニューロン間の結合重みを１にする。このように結合重みが変更されることで、詳細認識処理結果の転送が実現できる。これらの処理は、図１９において以下のように例示される。

【００７０】

１９１０から送られた詳細認識処理結果は、１９１１の指定部２３１から指定された頭部領域（１９１２）に送られる。送られた頭部に関する詳細認識処理結果は、ここでは、第１のＣＮＮの中間層特徴マップである１９０７の頭部領域（１９１２）に対してそのまま上書き処理を行うものとする。なお詳細は統合方法の説明において例示するが、第１のＣＮＮの中間層特徴マップと、第２のＣＮＮの出力する詳細認識処理結果と、の特徴量の次元数はここでは同じであるとする。最後に、詳細認識処理結果が統合された第１のＣＮＮの中間層出力は、実行部１２２第２層（図１９の１９１３）に送られ、以降も同様の統合処理を行ってもよいし、行わなくてもよい。第１のＣＮＮは、図１３で示したように、最終的に認識結果を出力する。以降も同様に統合処理を行う場合は、第２のＣＮＮのさらなる認識結果を得るために、１９０５をストリーム２２２３（頭部）第２層（図１９の１９１４）に送ってもよい。そして、第２のＣＮＮと第１のＣＮＮの認識結果の統合を行う際に、ある一つの層の特徴マップの出力を用いてもよいし、さらに深い層で得られた第２のＣＮＮの詳細認識処理結果を用いてもよい。なお、ここで述べる第２のＣＮＮの詳細認識処理結果とは、第２のＣＮＮの各層の中間層出力と、最終層出力とを含む。

【００７１】

今回の例では、統合する対象となる第２のＣＮＮの詳細認識処理結果と第１のＣＮＮの中間層出力は、同一の層の情報であるものとする。一方で、認識結果の統合には異なる層の出力を用いてもよい。例えば、第２のＣＮＮの第一層の詳細認識処理結果を、第１のＣＮＮの第三層の中間層特徴マップに統合してもよい。また、第２のＣＮＮの複数の異なる層の詳細認識処理結果を、第１のＣＮＮの特定の層の中間層特徴マップに統合してもよい。この場合の統合方法については後述する。

【００７２】

ステップＳ１８０８では、受信部２３３は、ＣＮＮ更新部１２３から誤差情報を受け取る。このとき受け取る誤差情報は、ステップＳ１８０７で送信した詳細認識処理結果に関

する誤差情報であり、ステップＳ１８０７で用いた注目領域情報に基づいて受信される。具体的には、ステップＳ１２０５で得られた第１のＣＮＮの誤差情報のうち、ステップＳ１８０７での送信先の領域に関わる誤差情報のみを受け取ることとなる。受信部２３３は、得られた誤差情報を更新部２２３に送る。更新部２２３は、受け取った誤差情報から、誤差逆伝播法に基づいて第２のＣＮＮ上の誤差を算出し、ここで得られた誤差情報を用いて第２のＣＮＮのモデルを更新する。このように第２のＣＮＮに逆伝播した誤差情報は、第２のＣＮＮに対して認識対象情報を入力した第１のＣＮＮの一部に再び伝播してもよく、第１のＣＮＮと第２のＣＮＮとでＥｎｄ－ｔｏ－ｅｎｄな学習系を構築することが可能である。ここで更新された第２のＣＮＮのモデルは、操作部２２２１に送られる。操作部２２２１は、ストリーム２２２３および統合部２２２５にモデルを送り、認識処理で用いるモデルの更新を行う。なお、ここでは変形部２２２４が変形処理に学習パラメータを用いていないものとして、変形部２２２４のモデルを更新しなかった。なお、変形処理になんらかの学習パラメータを用いている場合は、変形部２２２４のモデルを更新してもよい。なお、変形処理に関する詳細な説明は後述する。

【００７３】

そして、学習のためのMinibatchをすべて利用した場合には、処理はステップＳ１８０９を介してステップＳ１８１０に移行し、利用していないMinibatchが残っている場合には、処理はステップＳ１８０２に移行する。ここでの判定には、ステップＳ１２０７の判定結果をそのまま利用してもよい。

【００７４】

学習が完了した場合には、処理はステップＳ１８１０を介してステップＳ１８１１に移行する。一方、学習が完了していなければ、処理はステップＳ１８０２に移行する。ここでの判定には、ステップＳ１２０９の判定結果をそのまま利用してもよい。

【００７５】

ステップＳ１８１１では、操作部２２２１は、第２のＣＮＮのモデルを保存部２２４に送る。保存部２２４は、操作部２２２１から受けた第２のＣＮＮのモデルを記憶部Ｍ１に格納する。

【００７６】

以上、学習時に関する各部の詳細な構成および動作について説明を行った。次に、検出時に関する各部の詳細な構成および動作について説明を行う。前述のとおり、検出時には、図７のステップＳ７０１において第１のＣＮＮ及び第２のＣＮＮの認識処理を行う。ここで用いる各部の構成は、学習時と同じであってよい。すなわち、入力データを学習データではなく検出対象データとして、それ以外の部分を共通化して用いてよい。なお、モデルの更新などの学習処理は、検出時にも必要に応じて使い分けることができる。具体的には、検出対象データに教師データが付与されている場合、例えば学習時と同様の方法で学習を行うことによって、追加学習を行ってもよい。また、検出対象データに教師データが付与されていない場合、教師データに基づいて誤差を計算することはできないため、学習処理に関する一部の構成は用いなくてもよい。また、その場合、異常検知システムはユーザに対して教師データの付与を依頼してもよい。

【００７７】

以降では、詳細認識対象データをどのように設定するかに関して、一例を説明する。本実施形態においては、認識対象人物の行動（特に、キョロキョロという行動を異常として定めたときの行動）を認識するための構成の例を示した。ここでは、第１のＣＮＮの入力は認識対象人物の全身画像であり、第２のＣＮＮの入力は認識対象人物の頭部画像であるとしている。このうち、認識対象人物の頭部を第２のＣＮＮの詳細特徴抽出対象とする、ということが、詳細認識対象データを頭部と設定する、という定義に相当する。この定義は、“詳細認識対象データ＝頭部”という一対一の固定な関係を生じている。なお、詳細認識対象データを頭部以外に設定することも可能である。例えば、万引き行動を検知するために、“詳細認識対象データ＝商品に触れている手”などとしてもよい。この場合、例えば認識対象人物の手を検知することと、手が商品に触れていることを検知し、その手の

領域データを活用して詳細認識を行ってもよい。また、詳細認識対象データはユーザが指定してもよく、あらかじめシステム上に定義されていてもよい。

【００７８】

なお、詳細認識処理結果を統合するための一構成例として、本実施形態では、頭部の位置を考慮して第２のＣＮＮの詳細認識処理結果を第１のＣＮＮの中間層特徴マップに統合する例を示した。しかし、必要であれば、位置を考慮せず、第２のＣＮＮの詳細認識処理結果を第１のＣＮＮの中間層特徴マップに統合してもよい。その場合は、前出の変形処理を用いず、例えば第２のＣＮＮと第１のＣＮＮの各層の中間層出力を、例えば以下の文献４～６の方法で統合してもよい。この際、例えば第２のＣＮＮの詳細認識結果を、第１のＣＮＮの中間層特徴マップに統合する構成をとることができる。

【００７９】

文献４　…　Spatiotemporal Residual Networks for Video Action Recognition. C. Feichtenhofer, A. Pinz, R. P. Wildes. Advances in Neural Information Processing Systems (NIPS), 2016.

文献５　…　Spatiotemporal Multiplier Networks for Video Action Recognition. C. Feichtenhofer, A. Pinz, R. P. Wildes. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

文献６　…Convolutional Two-Stream Network Fusion for Video Action Recognition. C. Feichtenhofer, A. Pinz, A. Zisserman. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

また、頭部などの注目領域の位置を用いて、第２のＣＮＮの詳細認識処理結果と第１のＣＮＮの中間層特徴マップとを統合する方法として、その具体例を以下に述べる。なお以降で述べる方法は、これまで説明した「第２のＣＮＮの出力する詳細認識処理結果を第１のＣＮＮの特徴マップ上の領域に送信したうえで統合処理をする」という手順に必ずしも限らない点に注意が必要である。第一の候補としては、第２のＣＮＮの出力のみを用いる場合であり、例えば以下の式１を第１のＣＮＮの中間層特徴マップの注目領域に代入することができる。なお代入ではなく、足し合わせてもよい（加法演算）。

【００８０】

【数１】

$$f_1(\mathrm{Reshape}(A)) \cdots (1)$$

【００８１】

ここで、Aは第２のＣＮＮのある層の中間層出力ないし最終層出力（詳細認識処理結果）、Reshapeは後述の変形処理関数、f_1は特徴変換のための関数であり、例えば１×１畳み込み処理などを導入してもよい。また、f_1を用いない場合は恒等写像を用いてもよい。第二の候補としては、第２のＣＮＮの詳細認識処理結果と、第１のＣＮＮの注目領域に関する中間認識情報を用いる場合であり、例えば以下の式２を第１のＣＮＮの中間層特徴マップの注目領域に代入することができる。

【００８２】

【数２】

$$f_2(\mathrm{Reshape}(A), \mathrm{Crop}(B)) \cdots (2)$$

【００８３】

ここで、Bは第１のＣＮＮのある層の中間層出力、Cropは注目領域（頭部パーツの領域

）の特徴マップの出力を抜き出す処理であり、例えば以下の文献７のROI poolingを用いてもよい。f_2は特徴変換のための関数であり、Concatenate処理を行ったのち、１×１畳み込み処理を導入してもよい。これによって、第２のＣＮＮと第１のＣＮＮの頭部に関する局所的な中間層出力を考慮して認識処理を行うことができる。なお、f_2は、例えば文献８に記載のresidual networkを用いてもよい。

【００８４】

文献７　…　Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. S. Ren, K. He, R. Girshick, J. Sun. Advances in Neural Information Processing Systems (NIPS), 2015.

文献８　…　Identity Mappings in Deep Residual Networks. K. He, X. Zhang, S. Ren, J. Sun. ECCV, 2016

第三の候補としては、第２のＣＮＮの詳細認識処理結果と、第１のＣＮＮの中間層出力とを、一部の領域に限らずすべて用いる場合であり、例えば以下の式３を第１のＣＮＮの中間層特徴マップに代入することができる。

【００８５】

【数３】

$$f_3(\mathrm{PartialAdd}(\mathrm{Reshape}(A), B)) \cdots (3)$$

【００８６】

なお中間層特徴マップに代入するための構成は、ソフトウェアによる実現と、ハードウェアによる実現とがあり、図１８のステップＳ１８０７に係る説明で述べた方法によって実現可能なため、ここでは詳細な説明を省略する。ここで、PartialAddは第一引数を第二引数の一部の領域に足し合わせる関数であり、ここではBの注目領域に足し合わせる処理を担う。f_3は特徴変換のための関数であり、中間層特徴マップ全体の処理を行うことを特徴とする。そのため、より広い範囲の情報を捉えつつ、詳細認識処理結果との統合処理を行うことができるという特徴がある。第四の候補としては、第２のＣＮＮの詳細認識処理結果を、第１のＣＮＮの中間層特徴マップに基づいて入力する場合であり、例えば以下の式４を中間層特徴マップに代入することができる。

【００８７】

【数４】

$$f_5\left(\mathrm{PartialAdd}\left(\mathrm{Reshape}\left(A * \sigma(f_4(B))\right), B\right)\right) \cdots (4)$$

【００８８】

なお中間層特徴マップに代入するための構成は、ソフトウェアによる実現と、ハードウェアによる実現とがあり、図１８のステップＳ１８０７に係る説明で述べた方法によって実現可能なため、ここでは詳細な説明を省略する。ここでf_4、f_5は特徴変換の関数であり、σはsigmoid関数、*はelementwiseな掛け算処理である。すなわち、Bの関数f_4の出力の大きさに応じて、Aとの掛け算の結果が変化し、これによってAをどの程度重要な情報として扱うかの決定を行うことができる。なお、関数f_4をAにも依存するようにしてもよく、その場合は第２のＣＮＮと第１のＣＮＮの認識処理に基づいて統合処理をコントロールするシステムとなる。

【００８９】

なお、本実施形態では第２のＣＮＮの詳細認識処理結果を第１のＣＮＮの中間層特徴マップに統合する例を示した。しかし、逆に第１のＣＮＮの中間層特徴マップを第２のＣＮＮの詳細認識処理結果に統合し、その統合結果に基づいて第２のＣＮＮの処理を行ったのち、第１のＣＮＮの中間層特徴マップに統合するような処理を採用してもよい。

【００９０】

なお、本実施形態では第２のＣＮＮのある層の詳細認識処理結果を第１のＣＮＮに統合

する例を示した。その他の例として、第２のＣＮＮの複数の層の詳細認識処理結果を第１のＣＮＮに統合するような構成をとってもよい。この場合は、例えば複数層からなる詳細認識処理結果の大きさをすでに述べたReshapeによって揃え、Concatenateし、１×１畳み込み処理をして特徴次元数をもとの次元数に低減する。これにより、本実施形態ですでに述べた通常の詳細認識処理結果として扱うことができるため、このようにしてもよい。

【００９１】

なお、変形処理（上述のReshape関数）は、例えば、畳み込み処理に基づいて大きさを変更してもよいし、上記の非特許文献１で用いられるプーリング処理（max pooling, average poolingなど）を用いてダウンサンプリング処理を施してもよい。

【００９２】

なお、例えばmax poolingを用いる際に、入力特徴マップを均一にカバーしたプーリングを行うという指針のもとで、プーリング処理を行う窓のサイズを決める一つの例として、入力特徴マップサイズを出力特徴マップサイズで割ってもよい（例えばその値を天井関数で丸めてよい）。また、そこで決めた値のもとで、（入力マップサイズ－窓サイズ）／（出力特徴マップサイズ－１）という式によって、プーリング処理のストライドサイズを決めてもよい（その値を床関数で丸めてもよい）。なお出力特徴マップサイズは１以下である場合に、仮の値として２を用いることで、０割を回避することができる。また、窓のサイズはストライドの値以上の値であり、また入力特徴マップサイズと出力特徴マップサイズが同じ場合はストライドの値を１としてもよい。

【００９３】

なお、本実施形態では一枚のＲＧＢ画像を入力する第１のＣＮＮの例を示したが、他の入力形式であってもよい。例えば、以下の文献９で示されるTemporal Streamの入力層のように、複数の画像を入力として受け付ける入力層を持ってもよい。また、以下の文献９では、入力画像としてオプティカルフローを使う場合について例示されているが、本実施形態においても、そのようにしてもよいし、さらに文献９に記載のTwo-stream構成にしてもよい。また、文献１０のように、後段にLong-short term memoryなどのrecurrent neural network(RNN)を備える構成を用いてもよい。後段にLong-short term memoryなど時系列情報を処理する機能を持つコンポーネントを用いることで、時間情報の変化を捉えることが可能である。また、Long-short term memoryなどのRNNに畳み込み層を付与した公知のconvolutional long-short term memoryを用いてもよい。Long-short term memoryに畳み込み層を付与することで、後段の処理においても、本実施形態で用いる統合処理における第１・２のＣＮＮの位置関係を考慮することができ、これによって認識精度が向上する場合がある。

【００９４】

文献９　…　Two-Stream Convolutional Networks for Action Recognition in Videos. K. Simonyan, A. Zisserman. NIPS, 2014.

文献１０　…　A Multi-Stream Bi-Directional Recurrent Neural Network for Fine-Grained Action Detection. B. Singh, et al. CVPR, 2016.

なお本実施形態では、第１のＣＮＮと第２のＣＮＮとが異なる入力画像（224画素×224画素の画像サイズにリサイズ済み）を受け付け、第２のＣＮＮの詳細認識処理結果を第１のＣＮＮの中間層特徴マップに統合するマルチストリームの構成例を示している。しかし、上記の構成ではなく、第２のＣＮＮと第１のＣＮＮとを連携する他の構成を用いてもよい。例えば、第２のＣＮＮが第１のＣＮＮの中間層出力を受け取る構成や、第２のＣＮＮが入力画像を受け取らない構成や、入力画像をリサイズしない構成が考えられる。このようなアーキテクチャを組み合わせる例として上記すべての特徴を兼ねた構成の例を図２１に示す。

【００９５】

図２１の２１０１は、入力画像としての人体領域１０１１と同じものであり、第１のＣＮＮは人体領域１０１１と同じものである人体領域２１０１を受け取っている。図１９の全身画像１９０３とは異なり、人体領域２１０１はリサイズされていない。実行部１２２

第１層２１０２は人体領域２１０１を受け取り、畳み込み処理などをかけ、中間層出力２１０３として出力する。そして、中間層出力２１０３と、その頭部領域の中間層出力２１０４とは第２のＣＮＮに送出される。具体的には、中間層出力２１０４を第２のＣＮＮの入力として畳み込み処理などをストリーム２２２３第１層２１０６で行い、その出力である詳細認識処理結果と中間層出力２１０３とを統合部２２２５第１層２１０５で統合処理する。そして、その処理結果を詳細認識処理結果として、中間層出力２１０３に上書きする。そして、その結果を実行部１２２第２層２１０７に入力し、以降の処理を繰り返す。最後に、認識処理対象のクラス数の特徴次元数を出力する畳み込み処理を行い、特徴マップ全体を包含するMax pooling処理を行い、その結果をSoftmax処理にかけることで、認識結果を得ることができる。以上の統合処理の方法や、ニューラルネットワークのアーキテクチャなどは、（任意）に変更してもよい。これによって、図１９とは異なる構成で、第２のＣＮＮと第１のＣＮＮとを組み合わせることが可能である。

【００９６】

なお、本実施形態における説明では、例として詳細認識処理結果を第１のＣＮＮに統合する場合を示した。しかし、第１のＣＮＮではなく、Long-short term memory、ニューラルネットワーク、確率モデル、識別器、Convolutional long-short term memoryなどその他のモデルに統合するような構成であってもよい。このとき、第１のモデルがＣＮＮではなく、その他のモデルであってもよいし、ＣＮＮではあるが一部のみその他のモデルであってもよい（第２のＣＮＮも同様である）。

【００９７】

なお、本実施形態においてROI embeddingを行う場合の例を図２４に示す。図２４の２７０１は時系列を表している。２７０２および２７０３はそれぞれある時刻におけるある人物の画像および右手周辺の画像であり、それぞれ第２・第１のＣＮＮの中間層特徴マップを模擬的に示している。また、２７０４は第２のＣＮＮの中間層特徴マップをプーリングする際のプーリング範囲の例を示しており、２７０５は第１の中間層特徴マップ２７０３上の当該人物の右手周辺（注目）領域として、２７０２を統合処理２７０６によって統合することを示す。２７０７、２７０８はそれぞれ２７０２、２７０３と同様であるが、異なるタイミングで取得された特徴量である。２７１０は２７０６と領域のサイズおよび位置が異なることが分かる。この場合、２７０９のように、プーリング範囲を変更することで対応することが可能である。なお、２７０３などに示される画像は文献１２に記載のデータベースに収録された画像を使用している。

【００９８】

文献１２　…　A Database for Fine Grained Activity Detection of Cooking Activities. Marcus Rohrbach and et al., Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on.

なお、本実施形態において示される統合処理の効果や、中間処理結果などを可視化することができ、例えば図２７のように示してもよい。３００１、３００２はそれぞれ第１・２のＣＮＮの中間層特徴マップの模式図であり、３００３の示す注目領域に対して３００２を統合する３００４があるとする。このとき、例えば中間層特徴マップ（注目領域３００３などの局所領域でもよいし、全体でもよい）の活性具合（ノルムの大きさや数値の大きさ）や、その他の統計量などで示される数値が変化したことなどによって、統合処理の影響の大きさを測ることができる。このとき、上述の数値を例えばマップ上に示すことで、上記影響の大きさを可視化することができる。図２７では、３００５の部分の値の大きさが増えていることを例示している。なお、値の大きさを測る際には特定の次元の特徴マップを用いてもよいし、次元方向に総和などの処理を施して得た特徴マップを用いてもよい。

【００９９】

なお、本実施形態においては一例として第１・２のＣＮＮの学習手順などを示したが、例えば以下のような手順で学習をしてもよい。まず、第１・２のＣＮＮを独立に学習する。このとき、各ＣＮＮは例えば前出のVGG16を用いてもよい。次に、統合処理を行うレイ

ヤーを決定し、第１のＣＮＮを第２のＣＮＮと統合したアーキテクチャのもとで、前の学習処理で得た各ＣＮＮのパラメータを再利用する形で学習を行う（統合処理を行うレイヤーは何層あってもよく、例えば前述のVGG16の定義として4_2層および5_2層のReLU処理の結果を用いることができる）。このとき、第１・２のＣＮＮだけでなく、統合部分の学習処理を行ってよい。次に、Long-short term memoryの学習を行う。例えば上述の第１のＣＮＮの4_2層および5_2層において第２のＣＮＮと統合処理を行い、その結果をconcatenate処理したうえでベクトルとして扱い、Long-short term memoryの入力として用いてもよい。

【０１００】

以上説明したように、本実施形態で示した一形態は、データ中の認識対象人物やその状態を認識するＣＮＮと詳細ＣＮＮとがあり、ＣＮＮの中間層出力と詳細ＣＮＮの詳細認識処理結果とを統合するための学習を行う。これにより、認識に寄与する詳細な特徴量を抽出し、認識対象を精度よく認識することができ、最終的な認識結果（異常検知結果）の精度が向上しうる。

【０１０１】

［第２の実施形態］

本実施形態を含む以下の各実施形態では、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か居切は第１の実施形態と同様であるものとする。第１の実施形態では、データ中の物体やその状態を認識する畳み込みニューラルネットワーク（ＣＮＮ）である第１のＣＮＮと、さらに詳細な特徴を認識するＣＮＮである第２のＣＮＮと、があり、学習に基づいてそれらの中間の認識処理結果を統合して認識結果を得る例を示した。その際の構成の第一の例は、第１のＣＮＮと第２のＣＮＮとがそれぞれ一つあり、組として動作するTwo-streamの構成であった。また、その他の例として、第１のＣＮＮと第２のＣＮＮの組が複数ある場合の構成について例示した。これらはいずれも第１のＣＮＮと第２のＣＮＮとが組として一対一の対応関係にあるという特徴がある。

【０１０２】

本実施形態では、一つの第１のＣＮＮに対して複数の第２のＣＮＮが存在し、複数の第２のＣＮＮが異なる詳細認識処理結果を抽出する場合の構成および動作について例示する。このとき、複数の第２のＣＮＮの詳細認識処理結果をいかに統合するか、ということに関して、第１の実施形態とは異なる構成および動作について説明する。なお、本実施形態で示す構成は第１の実施形態で例示した構成と大部分が同一であり、一部の構成と動作が異なる。本実施形態において第１の実施形態と構成や動作が大きく異なるのは、先述のように、一つの第１のＣＮＮに対して複数の第２のＣＮＮが存在する点である。

【０１０３】

本実施形態においては、一つの構成例として、第１の実施形態の構成に加えて手の詳細な特徴量を抽出する第２のＣＮＮを導入する場合の構成および動作に関して例示する。すなわち、認識対象人物の全身を認識する第１のＣＮＮと、頭部を詳細に認識する第２のＣＮＮ（頭部）と、右手を詳細に認識する第２のＣＮＮ（右手）と、左手を詳細に認識する第２のＣＮＮ（左手）と、があるとする。これによって、例えば認識対象人物が万引き行動を行う際に、対象人物の全身と、頭部と、右手と、左手と、の詳細な特徴量を抽出・統合することで、万引き行動を比較的精度よく検知しうる。

【０１０４】

本実施形態において、認識対象人物の全身と、頭部と、右手と、左手と、の領域を抽出する方法は、第１の実施形態と同一の方法を用いるとする。ここで、図１０の下部における１０１１、１０１２、１０１３、１０１５はそれぞれ認識対象人物の全身、頭部、右手、左手、が検出された領域を例示する。このうち、全身および頭部領域中の特徴量は図１９の構成によって認識処理をかけることができ、図１９は第１のＣＮＮと第２のＣＮＮとが一対一の関係である場合の例を示している。このとき、図１９の頭部領域に関する処理をそれぞれ右手領域や左手領域に読み替えることで、「全身および右手領域」や、「全身および左手領域」の認識処理を行う構成になりうる。すなわち、複数の第２のＣＮＮを用

意し、それぞれが頭部・右手・左手に関する詳細認識処理結果を抽出する機能を持つものとすることができる。このとき重要になるのは、上記の複数の詳細認識処理結果をどのように統合するか、という点である。例えば、図１９の１９０９において、頭部・右手・左手に関する詳細認識処理結果の変形結果をそれぞれ得たとする。これら複数の変形結果と、第１のＣＮＮの中間層出力１９０７と、を統合処理する機能が必要である。これらの特徴量を統合する方法としては、例えば各パーツの位置を考慮しない方法として、以下の式５を第１のＣＮＮの中間層特徴マップの注目領域に代入することができる。

【０１０５】

【数５】

$$f_6\left(\mathrm{Concat}(B, A, C, D)\right) \cdots (5)$$

【０１０６】

なお中間層特徴マップに代入するための構成は、ソフトウェアによる実現と、ハードウェアによる実現とがあり、図１８のステップＳ１８０７に係る説明で述べた方法によって実現可能なため、ここでは詳細な説明を省略する。ここでB、A、C、Dはそれぞれ全身・頭部・右手・左手の詳細認識処理結果であり、ConcatはConcatenateの略で複数の入力を特徴次元方向に結合する処理であり、f_6は特徴変換処理を表す。例えばf_6が１×１の畳み込み処理であり、特徴次元数が相対的に小さくなる場合、式５はB、A、C、Dの入力を次元圧縮する処理を行うことになる。式５に関して、パーツの位置を考慮する場合は、例えば以下の式６を第１のＣＮＮの中間特徴マップの注目領域に代入することができる。

【０１０７】

【数６】

$$f_7\left(\mathrm{Concat}\left(B, \mathrm{Pad}(\mathrm{Reshape}(A)), \mathrm{Pad}(\mathrm{Reshape}(C)), \mathrm{Pad}(\mathrm{Reshape}(D))\right)\right) \cdots (6)$$

【０１０８】

なお中間層特徴マップに代入するための構成は、ソフトウェアによる実現と、ハードウェアによる実現とがあり、図１８のステップＳ１８０７に係る説明で述べた方法によって実現可能なため、ここでは詳細な説明を省略する。ここでPadは領域を考慮した０埋め処理である。すなわち、第１のＣＮＮの特徴マップ上における各パーツの領域の大きさと位置とを考慮した０埋め処理であり、その具体的な処理の例を図２０に示す。

【０１０９】

２３０１、２３０２、２３０３はそれぞれ、頭部、右手、左手の詳細認識処理結果である。これらをReshape処理によって、統合先の第１のＣＮＮの特徴マップ上の大きさに変換したものが、２３１１、２３１２、２３１３である。さらにこれらを、統合先の第１のＣＮＮの特徴マップ上のパーツ位置に配置し、不要な領域を０埋め（Pad処理）した詳細認識処理結果が２３２１、２３２２、２３２３である。これによって、第１のＣＮＮの特徴マップ上での領域の大きさと位置とを考慮した詳細認識処理結果を得ることができる。これらの出力とBをConcatし、特徴変換する処理が式６である。f_7は特徴変換処理を表す。

【０１１０】

なおConcat処理ではなく、第１の実施形態のPartialAdd処理を用いても、パーツ領域の大きさと場所とを考慮した処理を導入することが可能である。また、例えば第１の実施形態に記載のその他の処理を導入して、複数の詳細認識処理結果の統合処理を導入してもよいし、その他の方法を用いてもよい。例えば、以下の式７を第１のＣＮＮの中間特徴マップに代入することができる。

【０１１１】

【数７】

$$f_8\left(\mathrm{PartialAdd}\left(B, \mathrm{Reshape}(A), \mathrm{Reshape}(C), \mathrm{Reshape}(D)\right)\right) \cdots (7)$$

【０１１２】
なお中間層特徴マップに代入するための構成は、ソフトウェアによる実現と、ハードウェアによる実現とがあり、図１８のステップＳ１８０７に係る説明で述べた方法によって実現可能なため、ここでは詳細な説明を省略する。基本的に式７の結果は第１のＣＮＮの中間特徴マップ全体に代入すればよい。ここでPartialAddは第１の実施形態で説明したパーツの場所を考慮した統合処理関数であり、f_8は特徴変換処理を表す。式７のPartialAddは第１の実施形態とは異なり、複数パーツに対応したPartialAdd関数である。その処理内容は基本的に第１の実施形態と同じであり、変数A, C, Dがそれぞれ持つ頭部・右手・左手の詳細な特徴量を、B中の領域に対して、足し合わせる処理を示す。
【０１１３】
なお、複数の領域が重複する場所に関しては、例えば遮蔽されたパーツの領域に関する詳細認識処理結果は無視する（統合処理を行わない）という処理を導入してもよい。この際、パーツが遮蔽されていることを判定する方法として、例えばパーツの奥行に関する位置を取得して用いることで判定してもよい。
【０１１４】
なお、それぞれの詳細認識処理結果を抽出するか否かを判定し、不要な詳細認識処理結果は用いないという機能を導入してもよい。例えば遮蔽されているパーツや、パーツ領域の検出結果の信頼度が低いパーツを検出し、その場合に詳細認識処理結果を抽出しない、あるいは詳細認識処理を行わない、というように処理の変更をすることができる。
【０１１５】
なお、第１のＣＮＮの中間層出力と第２のＣＮＮの詳細認識処理結果とを統合することで、値のスケールに変化が起こるのを避けるために、正規化処理を導入してもよい。例えば統合処理を行った第１のＣＮＮと第２のＣＮＮの数を用いてもよい。具体的には、第１のＣＮＮの中間層特徴マップのある領域には、もとの中間層出力と、２つの第２のＣＮＮの詳細認識処理結果とが足し算によって統合された結果が代入されたとする。すなわち、３つの種類の値の和が得られている。
【０１１６】
なお中間層特徴マップに代入するための構成は、ソフトウェアによる実現と、ハードウェアによる実現とがあり、図１８のステップＳ１８０７に係る説明で述べた方法によって実現可能なため、ここでは詳細な説明を省略する。ここで、該当領域部分の値を３で割ることで、正規化を行ってもよい。また、統合処理を行った第１のＣＮＮと第２のＣＮＮの数は陽には用いず、例えば平均値を用いて正規化を行ってもよい。具体的には、第１のＣＮＮの各特徴マップの縦方向および／または横方向の平均値を用いて、特徴マップの値から引くことで、特徴マップの値のセンタリングを行ってもよい。このとき中心化行列を用いると、平易にセンタリングを行うことができる。いま、中間層特徴マップAの縦方向に関する平均をmean(A,1)とし、Aの中心化行列をHとすると、以下の式８，９からＨＡ及びＡＨを求めることができる。
【０１１７】

【数８】

$$HA = A - \mathrm{mean}(A, 1) \cdots (8)$$

$$AH = A - \mathrm{mean}(A, 2) \cdots (9)$$

【０１１８】
正規化としてこれらの処理のいずれかを用いてもよいし、両方を導入したHAHを用いてもよい。なお中心化行列はＨ＝Ｉ－１／ｎ　1_n　1_n^Tとして求めることができる。ここで、nは対称行列Aの一辺の大きさであり、Iはn×nのidentity matrix、1_nはnサイズの1ベクトル、Tは転置処理を示す。これらの式は微分可能であり、誤差逆伝播法に用いることができる。なお、特徴マップが対称行列でない場合も同様にセンタリング処理を用いることができる。その他の正規化方法として、例えば以下の文献１１の方法を用いてもよいし、それ以外の方法を用いてもよい。
【０１１９】
文献１１　…　Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. S. Ioffe, C. Szegedy, JMLR.
以上のような構成においても、第１の実施形態で示したように、例えば誤差逆伝播法を用いて第１のＣＮＮと複数の第２のＣＮＮとを学習することが可能である。なお、第１の実施形態において図２１に関して示したような構成を用いて、複数の第２のＣＮＮの導入を図ってもよいし、その他の構成によって第１のＣＮＮと複数の第２のＣＮＮとの統合処理を行ってもよい。
【０１２０】
なお、本実施形態では第１のＣＮＮが一つであり、第２のＣＮＮが複数ある場合の例を示したが、第１のＣＮＮが複数存在する構成であってもよい。その構成における一つの例としては、次のようなものがある。ＲＧＢ画像を入力データとする第１のＣＮＮのストリーム、複数枚のオプティカルフローを入力データとするストリームの第１のＣＮＮのストリーム、を用意し、それぞれの第１のＣＮＮに関する詳細認識処理結果を抽出する第２のＣＮＮを複数用意してもよい。それ以外の構成としても、例えば第１の実施形態で示した構成を基本として、複数の第２のＣＮＮを導入する形をとってもよい。
【０１２１】
なお、本実施形態では複数の第２のＣＮＮをそれぞれ別個に保持する場合の例を示したが、第２のＣＮＮのモデルを一部ないし全部共通化してもよい。例えば、一般的な行動認識処理をするにあたっては、認識対象人物の右手の動作も左手の動作も意味合いは同じであり、それぞれの特徴を抽出するための構成は右手と左手とで共通であってもよい場合がある。そこで、認識対象人物の右手に関する詳細な特徴を抽出する第２のＣＮＮと、左手に関する第２のＣＮＮと、では、共通のモデルを用いることとしてもよい。また、必要であれば、複数の第２のＣＮＮのモデルをすべて共通化してもよい。これにより、大幅なモデルサイズ削減が期待できる。
【０１２２】
なお、本実施形態では詳細認識に用いる認識対象人物のパーツ（頭部や、手など）をあらかじめ詳細認識対象として設定し、それらをすべて認識処理に用いる場合の例を示した。このうち、認識に寄与する詳細認識対象を選択し、選択された詳細認識対象のみを認識処理に用いてもよい。例えば、式（４）は認識に寄与する詳細認識対象の特徴量を検出時に動的に重みづけする機能を持つ。その他の方法としては、例えば学習時に選択処理を行う場合がありえる。例えば、クロスバリデーションによって、どの詳細認識処理結果を用

いると認識処理が向上するかを確かめることが可能な場合は、これを用いてもよい。また、どの詳細認識処理結果を用いるかユーザに決定させてもよいし、必要であればユーザに新たな詳細認識対象を追加させてもよい。例えば、不審者を検知する不審行動検知のユースケースにおいては、人体のどの部分に不審さが現れるか、ということは現場の監視員が知悉するところである。そこで、監視員に認識対象人物のどのパーツに着目すべきかどうかという観点において、詳細認識対象をリコメンドしてもらうことが考えられる。検出時の実運用において新たな詳細認識対象が追加された場合は、必要であれば、はじめから学習処理をやり直してもよいし、使用していたモデルを初期値として学習処理をやり直してもよい。新たな詳細認識対象を検出するための一構成として、可能であれば上記の文献２の検出結果を用いてもよいし、必要であれば専用の検出器を導入してもよい。

【０１２３】

なお、本実施形態では詳細認識対象の種類が固定であるとしたが、動的に変動してもよい。例えば、認識対象の状態を認識するために、認識対象に付属する物体または物品の種類が重要である場合がありえる。具体的には、認識対象人物の属性を認識するために、当該人物の身に着けている品の種類が重要であるとする。しかしながら、検出時にどのような人物が現れるか事前には分からない場合、現れた人物が身に着けているものに応じて詳細に見るべき品の種類が変化しうる。これが、詳細認識対象の種類が動的に変動する場合に対応するとする。このような場合、物体／物品らしいものを検知することができる検出器を用いて、詳細認識対象の検知器として利用し、検知された詳細認識対象を認識処理に用いる構成を用いてもよい。例えば、上記の文献７の方法を用いて、認識対象人物が持つ物品／物体らしいものを検出し、これを詳細認識する構成が考えられる。その際、例えば検知された物品／物体がネックレスであるなら、ネックレス専用の第２のＣＮＮを用いて詳細認識処理をするように、物品／物体のおおまかな種類に応じた詳細認識処理の構成が考えられる。この際、どのような種類か不明な物品／物体であれば、種別不明品を対象として第２のＣＮＮを用いてもよい。また、物品／物体の種類によらずに、同一の第２のＣＮＮを用いて認識処理をすることが可能である。

【０１２４】

なお、本実施形態では、認識処理を行う例として認識対象人物に関する異常な行動を認識する処理や、一般的な行動を認識する例を示したが、これらは一例であって、その他の認識処理を目的とした場合にも上述の方法を用いてもよい。例えば、静止画像中のシーンを認識するタスクにおいて、詳細認識処理を行う第２のＣＮＮと、全体の荒い認識処理を行う第１のＣＮＮと、を用い、それらを統合する構成を用いてもよい。より具体的には、先述した認識対象人物の全身の画像を入力とする第１のＣＮＮと、認識対象人物のパーツに関する認識結果を出力する第２のＣＮＮと、に関する構成と同様に、シーンを認識すべき対象である全体の静止画像を入力とする第１のＣＮＮと、全体の画像の一部に含まれる物体の詳細な特徴量を抽出する第２のＣＮＮと、があり、これらの中間層出力と認識処理結果とを統合したうえで、シーンを認識する構成が考えられる。この際、例えば物体を検知する方法としては、上記の文献７の方法を用いてもよい。これらの具体的な処理の方法は、本実施形態ですでに述べた認識対象人物の行動認識に関する構成・動作例によって説明ができるため、詳細な処理の説明を割愛する。

【０１２５】

なお、統合処理の一例として、図２５を用いてもよい。２８０１は第１のＣＮＮのl層が中間層特徴マップを出力する例を示し（例えばl層におけるReLU処理後の中間層特徴マップ）、２８０２～２８０４はそれぞれ第２のＣＮＮ（注視領域は異なってよい）のl層が中間層特徴マップを出力する例を示している。このとき、第１のＣＮＮの中間層特徴マップに、第２のＣＮＮの中間層特徴マップを統合するとする。まず２８０５において、中間層特徴マップ２８０１を受け取る（２８１４を伝達して送信される）。ここでは、中間層特徴マップ２８０１に対して中間層特徴マップ２８０２をROI embeddingする。次に、２８０８で非線形処理をかける。このとき非線形処理として、例えばBN-ReLU-Convなど、文献４などで記載の方法を用いてもよい（なにも処理をしなくてもよい。例えば、ここで

は２８０１の出力を足し合わせる場合を例示したが、０で埋められた仮の特徴マップにROI embeddingしてもよいし、そのあとに非線形処理をかけなくてもよい）。２８０３、２８０４は２８０２と同様であり、２８０６、２８０７は２８０５と同様であり、２８０９、２８１０は２８０８と同様であり、２８１５、２８１６は２８１４と同様である。また、３つ以上の第２のＣＮＮが存在する場合を想定して、２８１８として例示した。これらの出力（中間層特徴マップの統合処理における中間状態）と、さらに２８０１の中間層特徴マップを、２８１１で統合する。このときの統合は、例えば足し算を用いてもよい（ここでは次元数は同じとする。次元数が異なる場合は別記参照のこと）。上記の足し算された結果が、２８１２で非線形処理にかけられる（ここでは非線形処理をかける例を示したが、かけなくてもよいし、必要であれば線形処理でもよい）。その後、２８１７でReLU処理にかけられ、２８１３で中間層特徴マップ２８０１とさらに足しあわされたうえで、次の第１のＣＮＮの中間層への入力として用いられる。なお、上記の構成は、文献４で示されるようなResidual Networkをマルチパーツストリームに拡張した例として考えることができる。

【０１２６】

なお、全体の統合処理アーキテクチャの例として、図２６を示す。２９０１～２９０３は、それぞれ認識対象人物、当該人物の左手、当該人物の右手の画像を示し、それぞれ第１のＣＮＮ２９０４（全身）、第２のＣＮＮ２９０５（左手）・２９０６（右手）に入力されるものとする。このとき、２９０８によってＣＮＮストリーム（左手）の中間層特徴マップが第１のＣＮＮ（全身）の中間層の統合処理部に送られることで統合され、また２９０５によって同様に第２のＣＮＮストリーム（右手）の中間層特徴マップが第１のＣＮＮに送られ統合される。図２６では、このような処理を複数の層で行うことができることを例示している。これらの統合結果が、２９１０によってＲＮＮ２９０７に送られる。ＲＮＮはLong-short term memoryなど、その他の公知のモデルでもよい。ＲＮＮを導入することで、このように、本実施形態におけるアーキテクチャをより長期の時間方向に対して拡張することができる。なお、２９１１は第２のＣＮＮストリームがさらにあってもよいことを示している。

【０１２７】

なお、全体の統合処理アーキテクチャの例として、図２８を示す。図２８の第１のＣＮＮ３１２０は、本実施形態に係る第１のＣＮＮのネットワーク構造の一例を示したものである。図２８に示す如く、第１のＣＮＮ３１２０は、入力層３１０１、ｃｏｎｖｏｌｕｔｉｏｎ１層３１０２、ｐｏｏｌｉｎｇ１層３１０３、ｃｏｎｖｏｌｕｔｉｏｎ２層３１０４、ｐｏｏｌｉｎｇ２層３１０５、Ｉｎｎｅｒｐｒｏｄｕｃｔ１層３１０６、Ｉｎｎｅｒｐｒｏｄｕｃｔ２層３１０７、出力層３１０８、を有している。また、２つの階層間の処理方法として、ｃｏｎｖｏｌｕｔｉｏｎ処理３１１０、ｐｏｏｌｉｎｇ処理３１１１、ＩｎｎｅｒＰｒｏｄｕｃｔ処理３１１２が設定されている。それぞれの処理の具体的な内容は周知であり、例えば図１３と同様であるため、ここでは省略する。また、第２のＣＮＮ３２２０は構成としては第１のＣＮＮとここでは同じであるとし（違ってもよい）、３２０２～３２１２はそれぞれ３１０２～３１１２に対応しているとする。このとき、第２のＣＮＮの１層目の出力（ｃｏｎｖｏｌｕｔｉｏｎ１層３２０２がかけられた後や、さらにＲｅＬＵをかけた後でもよい）を統合処理する例を３３０１に示す。３３０１は第２のＣＮＮの中間層特徴マップから第１のＣＮＮの中間層特徴マップに送信され、統合処理が行われる例である。

【０１２８】

以上説明したように、本実施形態で示した一形態は、データ中の認識対象人物やその状態を認識するＣＮＮと詳細ＣＮＮとがあり、ＣＮＮの中間層出力と詳細ＣＮＮの認識処理結果とを統合するための学習を行う。このとき、一つのＣＮＮに対して複数の詳細ＣＮＮがあり、複数の認識処理結果を統合する処理を行う。これにより、認識対象を精度よく認識するための複数の詳細な特徴量に基づいて認識処理を行うことができ、最終的な認識結果（異常検知結果）の精度が向上しうる。

【０１２９】
［第３の実施形態］
第１，２の実施形態では、第１のＣＮＮの中間層出力と、第２のＣＮＮの詳細認識処理結果と、を統合することによって、単一の目的における学習・認識処理を行う例を示した。
【０１３０】
本実施形態では、上記の例に加えて、複数の目的関数を用いて、複数の目的における学習・認識処理を行う例を示す。具体的には、学習時に第１のＣＮＮと第２のＣＮＮとで異なる目的関数を持つマルチタスク学習を行う場合の例を示す。
【０１３１】
なお、本実施形態で示す構成は第１，２の実施形態で例示した構成と大部分が同一であり、一部の構成と動作が異なる。本実施形態において第１，２の実施形態と構成や動作が大きく異なるのは、先述のように、第１のＣＮＮと一つ以上の第２のＣＮＮとがある構成において、マルチタスク学習を行う点である。
【０１３２】
本実施形態においては、一つの構成例として、第１の実施形態の構成においてマルチタスク学習を行う場合の構成および動作に関して例示する。すなわち、認識対象人物の全身の特徴量を抽出する第１のＣＮＮと、頭部の詳細な特徴量を抽出する第２のＣＮＮと、があり、第２のＣＮＮは第１のＣＮＮとは異なる目的関数に基づいて学習処理を行うとする。これによって、例えば全身の行動と頭部の行動とで、異なる意味を持つ行動を別々に識別処理するための学習が可能になり、また、それぞれのタスクに必要な特徴量を内部的に統合する処理がなされることで、最終の認識処理の精度が向上することが期待できる。
【０１３３】
全身の認識処理を行う第１のＣＮＮの目的関数と、頭部の詳細認識処理を行う第２のＣＮＮの目的関数と、を以下に式１０，１１として示す。
【０１３４】
【数９】

$$\min Loss_{\mathrm{Body}} = \Sigma\left(t_{\mathrm{Body}} - y_{\mathrm{Body}}\right)^2 \cdots\cdots(10)$$

$$\min Loss_{\mathrm{Head}} = \Sigma\left(t_{\mathrm{Head}} - y_{\mathrm{Head}}\right)^2 \cdots\cdots(11)$$

【０１３５】
ここで、$Loss_{Body}$、$Loss_{Head}$はそれぞれ、認識対象人物の全身に関する認識処理を行う第１のＣＮＮのロス関数、認識対象人物の頭部に関する詳細認識処理を行う第２のＣＮＮのロス関数である。t_{Body}、t_{Head}、y_{Body}、y_{Head}はそれぞれ前出の第１のＣＮＮの教師データ、第２のＣＮＮの教師データ、第１のＣＮＮの認識結果、第２のＣＮＮの詳細認識処理結果を表す。なお、式（１０）、（１１）のΣはMinibatchに含まれるデータ分だけ総和を得て、平均値を得る処理を行う関数とする。これらのロス関数は平均二乗誤差と呼ばれる。なお、平均二乗誤差ではなく、その他のロス関数を用いてもよい。このとき問題になるのは、上記の教師データをどのようなものにするかと、認識結果をどのように得るか（ネットワークアーキテクチャの問題）、ということである。
【０１３６】
第１のＣＮＮの教師データおよび認識結果は、第１の実施形態で示した方法によって得ることができる。第２のＣＮＮの教師データおよび認識結果は、第１の実施形態では用いなかったので、ここで図１および図２２に基づいて新たに説明を行う。
【０１３７】

まず、第２のＣＮＮの教師データは、第１のＣＮＮの教師データと同様に図１の記憶部Ｄ１から読み込まれ、前段処理部１１から注目領域制御部２１に送信され、注目領域制御部２１から認識部２２ａに送信されるものとする。ここで認識部２２ａは、第１の実施形態で用いる図１の認識部２２の代わりに本実施形態で用いる機能であり、その構成の例を図２２に示す。

【０１３８】

実行部２２２ａは、実行部２２２とは異なり、注目領域制御部２１から送られてきた入力データに対する詳細認識処理結果（具体例：ラベルの推定結果）と、第２のＣＮＮの教師データとを、更新部２２３ａに送る。更新部２２３ａは、受け取った詳細認識処理結果と、第２のＣＮＮの教師データとを用いて、誤差を算出する。ここで用いる誤差算出方法は、第１の実施形態で示した方法を用いてよいし、式１１で例示した方法を用いてもよい。これらの誤差算出方法によって得られた誤差情報は、第１の実施形態で説明した方法によって学習に用いることができる。

【０１３９】

もし、式１０，１１の２種の教師データが完全に同一のものであれば、式１０，１１は異なる目的関数ではなく、同一の目的を持つ二つの目的関数となる。そのように学習処理を行ってもよいが、本実施形態では前述のとおり、異なる目的関数を持つ場合について説明を行うために、２種の教師データが異なる値を持ちうるという仮定のもとで説明を行う。例えば、第１のＣＮＮは認識対象人物の状態が不審であるか否かを判定し、第２のＣＮＮは認識対象人物が「キョロキョロ」行動をしたか否かを判定する、とする。このとき、式１０，１１は異なる目的のもとで認識結果を評価するものの、「不審さ」と「キョロキョロ行動」は比較的相関の高い状態であると考えることができる。すなわち、式１０，１１に基づくマルチタスク学習は、二つの目的に共通して有用な特徴を抽出するための機構を提供することが可能である。

【０１４０】

なお、上記のような教師データの組だけでなく、その他の教師データを用いて学習処理を行ってもよい。どのような教師データの組を用いるかは、解くべき問題に依存するため、検出の目的に応じて適宜変えてもよい。

【０１４１】

マルチタスク学習のためのネットワークアーキテクチャの構成方法として、例えば以下のような方法を用いてもよい。図１９の上部は頭部に関する第２のＣＮＮのストリーム、下部は全身に関する第１のＣＮＮのストリームに関する構成の例を表している。このとき、上部のストリームの出力が、式１１に基づいて評価され、下部のストリームの出力が、式１０に基づいて評価されるとする。このような場合、第１のＣＮＮと第２のＣＮＮとで、それぞれ式１０，１１の異なる目的関数を用いる構成になり、すなわちマルチタスク学習の動作が可能な構成になる。なお、検出時には、目的に応じて、第１のＣＮＮと第２のＣＮＮの出力を用いてよい。具体的には、検出時に認識対象人物の「不審さ」を出力すべきであるなら第１のＣＮＮの出力を用いることにしてよく、また「キョロキョロ行動」をしたか否かを出力する必要があれば、第２のＣＮＮの出力を用いることにしてもよく、また必要であれば両方の出力結果を用いてもよい。なお、上述の方法以外のマルチタスク学習の構成例を用いて、第１のＣＮＮと第２のＣＮＮのマルチタスク学習を行ってもよい。

【０１４２】

以上説明したように、本実施形態で示した一形態は、データ中の認識対象人物やその状態を認識するＣＮＮと詳細ＣＮＮとがあり、ＣＮＮの中間層出力と詳細ＣＮＮの詳細認識処理結果とを統合するための学習を行う。このとき、ＣＮＮと（複数の）詳細ＣＮＮとで、それぞれ異なる目的関数を用いる場合について示した。これにより、認識対象を精度よく認識するための複数の詳細な特徴量に基づいて認識処理を行うことができ、最終的な認識結果の精度が向上しうる。

【０１４３】

［第４の実施形態］

図１～４，８，１１，１４～１７，２２に示した各機能部はハードウェアで実装しても良いが、一部をソフトウェア（コンピュータプログラム）で実装しても良い。例えば、記憶部として説明した機能部をメモリで実装し、それ以外の機能部をソフトウェアで実装しても良い。後者の場合、このようなソフトウェアを実行可能なコンピュータ装置は上記の異常検知システムに適用可能である。

【０１４４】

異常検知システムに適用可能なコンピュータ装置のハードウェア構成例について、図２３のブロック図を用いて説明する。なお、図２３に示したハードウェア構成は、以上説明した異常検知システムに適用可能なコンピュータ装置のハードウェア構成の一例に過ぎず、その構成は適宜変更しても構わない。

【０１４５】

なお、以上説明した異常検知システムは、認識装置１０、詳細認識装置２０、端末装置４０、それ以外の記憶部等の各装置を別個の装置としても良いし、一部の装置を一体化させても良い。然るに図２３に示した構成を有するコンピュータ装置は、認識装置１０、詳細認識装置２０、端末装置４０のそれぞれ若しくはその一部に適用しても良いし、異常検知システムを構成する装置のうち一部を一体化させた装置に適用しても良い。なお、以上説明した異常検知システムをどのような装置で構成するのかについては、適宜変形例が考えられる。

【０１４６】

ＣＰＵ２６０１は、ＲＡＭ２６０２やＲＯＭ２６０３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処理を実行する。これによりＣＰＵ２６０１は、コンピュータ装置全体の動作制御を行うと共に、該コンピュータ装置を適用した装置が行うものとして上述した各処理を実行若しくは制御する。

【０１４７】

ＲＡＭ２６０２は、ＲＯＭ２６０３や外部記憶装置２６０６からロードされたコンピュータプログラムやデータ、Ｉ／Ｆ２６０７を介して外部から受信したデータ、等を格納するためのエリアを有する。さらにＲＡＭ２６０２は、ＣＰＵ２６０１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ＲＡＭ２６０２は、各種のエリアを適宜提供することができる。ＲＯＭ２６０３には、コンピュータ装置の基本プログラムや設定データなど、書換不要のコンピュータプログラムやデータが保存されている。

【０１４８】

操作部２６０４は、キーボードやマウスなどのユーザインターフェースにより構成されており、ユーザが操作することで各種の指示をＣＰＵ２６０１に対して入力することができる。

【０１４９】

表示部２６０５は、ＣＲＴや液晶画面などにより構成されており、ＣＰＵ２６０１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、操作部２６０４と表示部２６０５とを一体化させてタッチパネル画面を構成しても良い。

【０１５０】

外部記憶装置２６０６は、ハードディスクドライブ装置等の大容量情報記憶装置である。外部記憶装置２６０６には、ＯＳ（オペレーティングシステム）や、本コンピュータ装置を適用した装置が行うものとして上述した各処理をＣＰＵ２６０１に実行させるためのコンピュータプログラムやデータが保存されている。外部記憶装置２６０６に保存されているコンピュータプログラムには、本コンピュータ装置を適用した装置の各機能部の機能をＣＰＵ２６０１に実行させるためのコンピュータプログラムが含まれている。また、外部記憶装置２６０６に保存されているデータには、上記の説明において既知の情報として説明したもの（パラメータや閾値、関数など）が含まれている。外部記憶装置２６０６に保存されているコンピュータプログラムやデータは、ＣＰＵ２６０１による制御に従って適宜ＲＡＭ２６０２にロードされ、ＣＰＵ２６０１による処理対象となる。

【０１５１】

Ｉ／Ｆ２６０７は、外部の機器との間のデータ通信を行うためのインターフェースとして機能するものである。ＣＰＵ２６０１、ＲＡＭ２６０２、ＲＯＭ２６０３、操作部２６０４、表示部２６０５、外部記憶装置２６０６、Ｉ／Ｆ２６０７は何れもバス２６０８に接続されている。

【０１５２】

なお、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し、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。また、以上の説明において使用した数値や構成や処理順は説明上一例としてあげたものであり、以上の説明に限らない。

【０１５３】

なお、以上説明した異常検知システムは、次のような各部を有する構成を有する情報処理装置の一例として説明したものである。

【０１５４】

・　認識対象の情報を含む認識対象データを入力として畳み込みニューラルネットワークの認識処理の結果を取得する認識処理部

・　認識対象データおよび／または畳み込みニューラルネットワークの中間層出力に対する注目領域を設定する注目領域設定部

・　注目領域中の認識対象データおよび／または中間層出力に対して認識処理よりも詳細な詳細認識処理を行う詳細認識部

・　詳細認識処理の結果と、中間層出力と、を統合処理する統合処理部

・　統合処理の結果を中間層出力として畳み込みニューラルネットワークに入力する中間入力処理部

・　認識処理の結果を出力する出力部

また、以上説明した異常検知システムは、次のような各部を有する構成を有する情報処理装置の一例として説明したものである。

【０１５５】

・　学習対象の情報を含む学習データを入力として畳み込みニューラルネットワークの認識処理の結果を取得する認識処理部

・　学習データおよび／または畳み込みニューラルネットワークの中間層出力に対する注目領域を設定する注目領域設定部

・　注目領域中の学習データおよび／または中間層出力に対して認識処理よりも詳細な詳細認識処理を行う詳細認識部

・　詳細認識処理の結果と、中間層出力と、を統合処理する統合処理部

・　統合処理の結果を中間層出力として畳み込みニューラルネットワークに入力する中間入力処理部

・　認識処理部、注目領域設定部、詳細認識部、統合処理部、中間入力処理部、のいずれか一つ以上に関する学習処理を行う学習部

（その他の実施例）

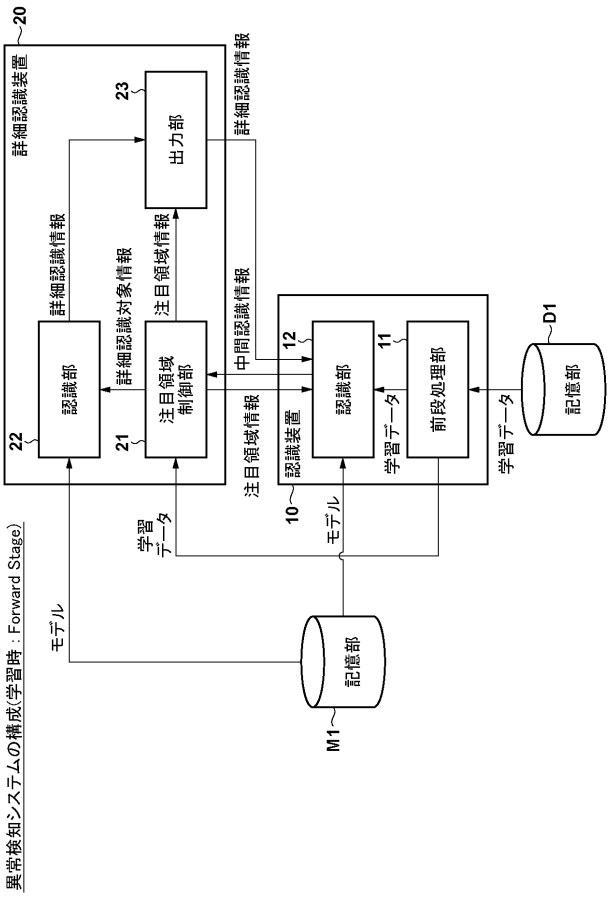
本発明は、上述の実施形態の１以上の機能を実現するプログラムを、ネットワーク又は記憶媒体を介してシステム又は装置に供給し、そのシステム又は装置のコンピュータにおける１つ以上のプロセッサがプログラムを読出し実行する処理でも実現可能である。また、１以上の機能を実現する回路（例えば、ＡＳＩＣ）によっても実現可能である。

【符号の説明】

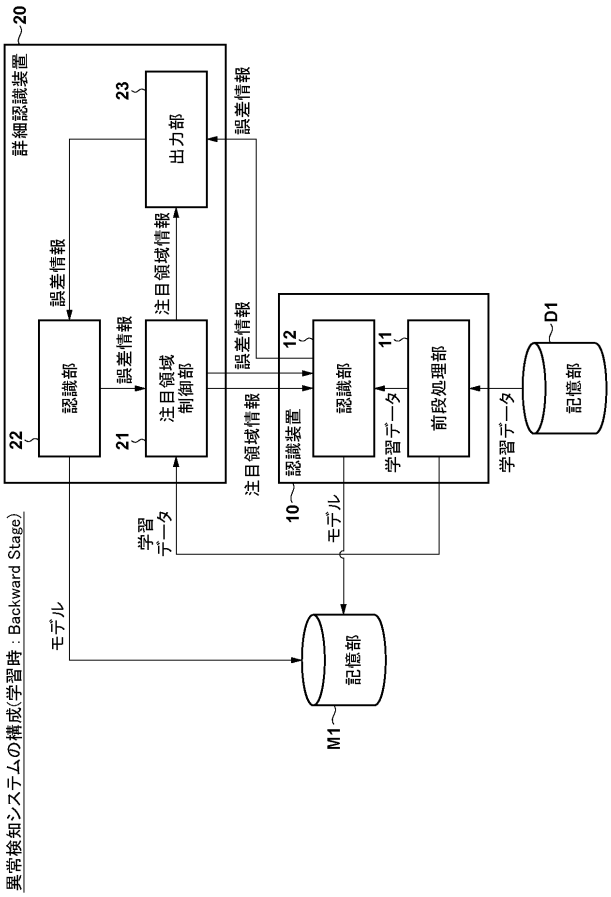
【０１５６】

１０：認証装置　２０：詳細認識装置　Ｄ１：記憶部　Ｍ１：記憶部

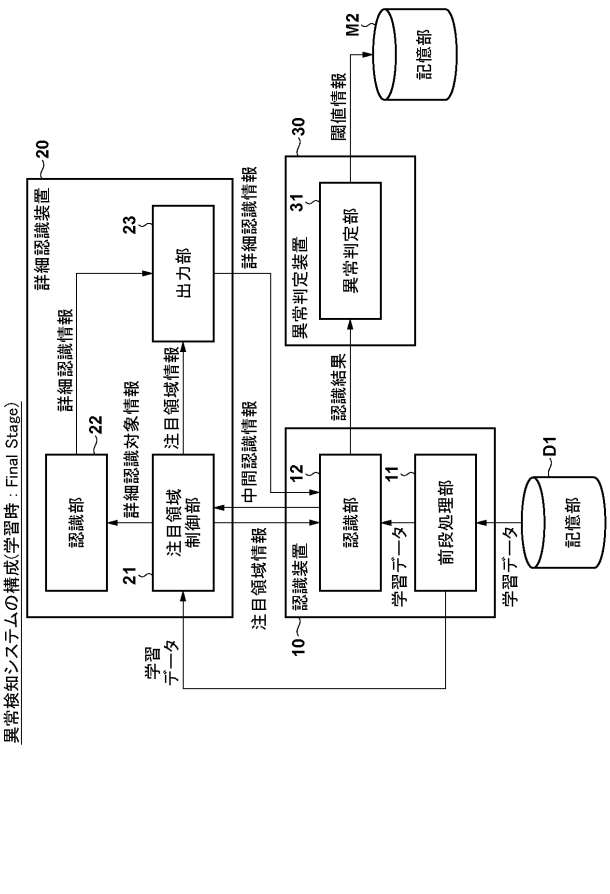
【図１】


異常検知システムの構成(学習時：Forward Stage)
詳細認識装置
20
認識部
22
詳細認識情報
注目領域制御部
21
詳細認識対象情報
出力部
23
注目領域情報
詳細認識情報
中間認識情報
認識装置
10
認識部
12
前段処理部
11
学習データ
注目領域情報
学習データ
D1
記憶部
モデル
学習データ
モデル
M1
記憶部

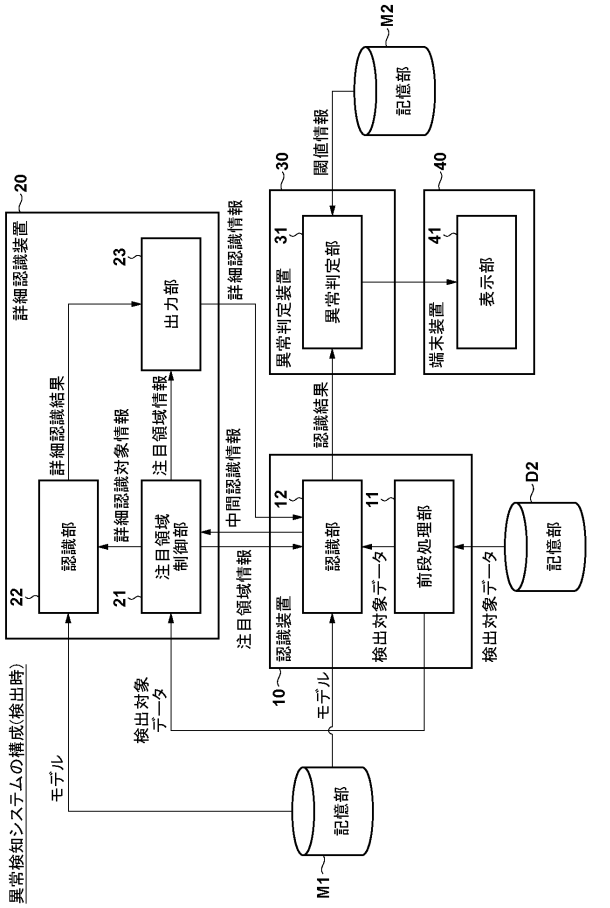
【図２】


異常検知システムの構成(学習時：Backward Stage)
詳細認識装置
20
認識部
22
誤差情報
注目領域制御部
21
誤差情報
出力部
23
注目領域情報
誤差情報
認識装置
10
認識部
12
前段処理部
11
学習データ
注目領域情報
学習データ
D1
記憶部
モデル
学習データ
モデル
M1
記憶部

【図３】

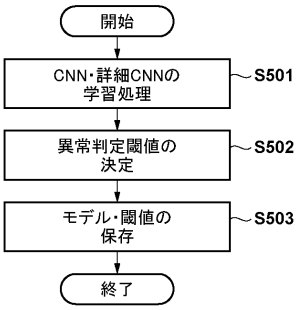

異常検知システムの構成(学習時：Final Stage)
詳細認識装置
20
認識部
22
詳細認識情報
注目領域制御部
21
詳細認識対象情報
出力部
23
注目領域情報
詳細認識情報
中間認識情報
認識装置
10
認識部
12
前段処理部
11
学習データ
認識結果
異常判定装置
30
異常判定部
31
閾値情報
記憶部
M2
注目領域情報
学習データ
学習データ
D1
記憶部
学習データ

【図４】


異常検知システムの構成(検出時)
詳細認識装置
20
認識部
22
詳細認識結果
注目領域制御部
21
詳細認識対象情報
出力部
23
注目領域情報
詳細認識情報
中間認識情報
認識装置
10
認識部
12
前段処理部
11
検出対象データ
認識結果
異常判定装置
30
異常判定部
31
閾値情報
記憶部
M2
端末装置
40
表示部
41
検出対象データ
検出対象データ
D2
記憶部
モデル
検出対象データ
モデル
M1
記憶部

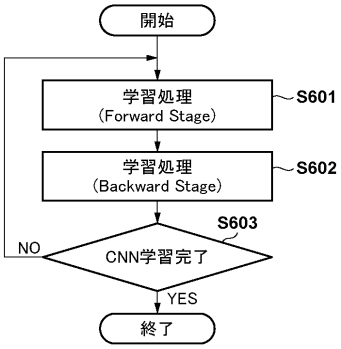
【図５】

異常検知システムの動作(学習時)



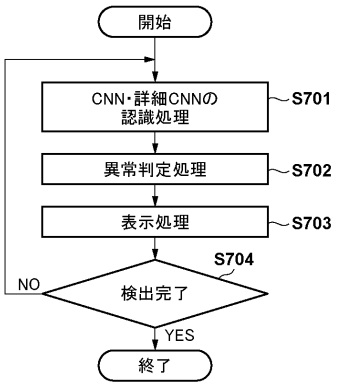
【図６】

ステップS501の詳細動作



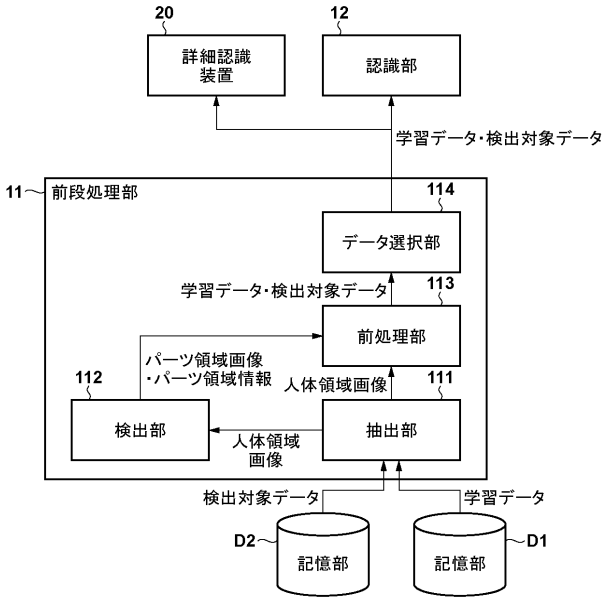
【図７】

異常検知システムの動作(検出時)



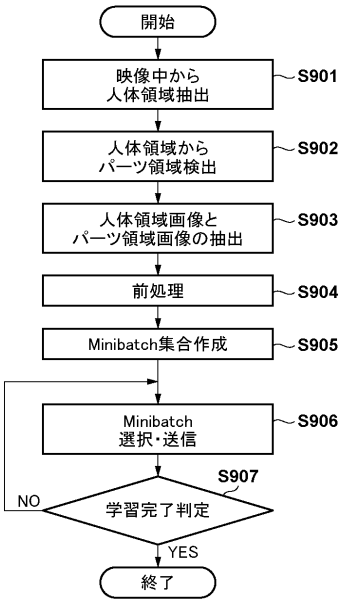
【図８】

前段処理部11の構成(学習時・検出時併記)

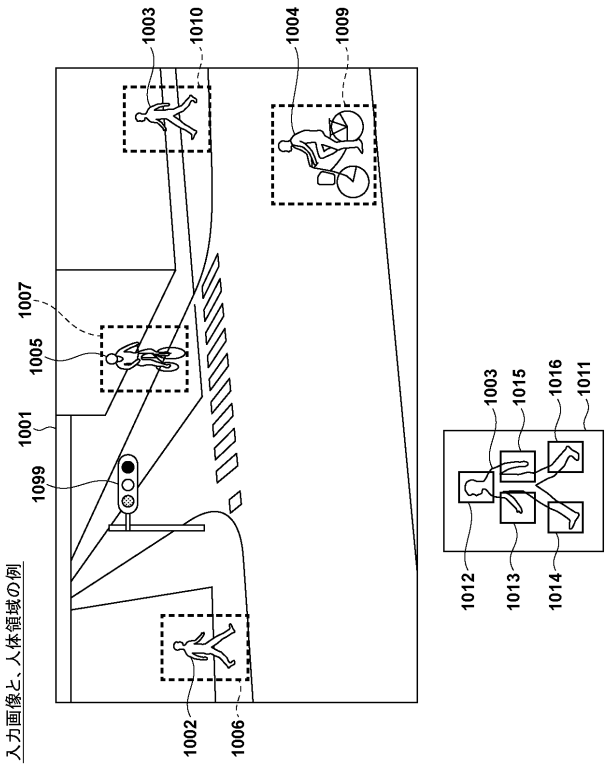


【図９】

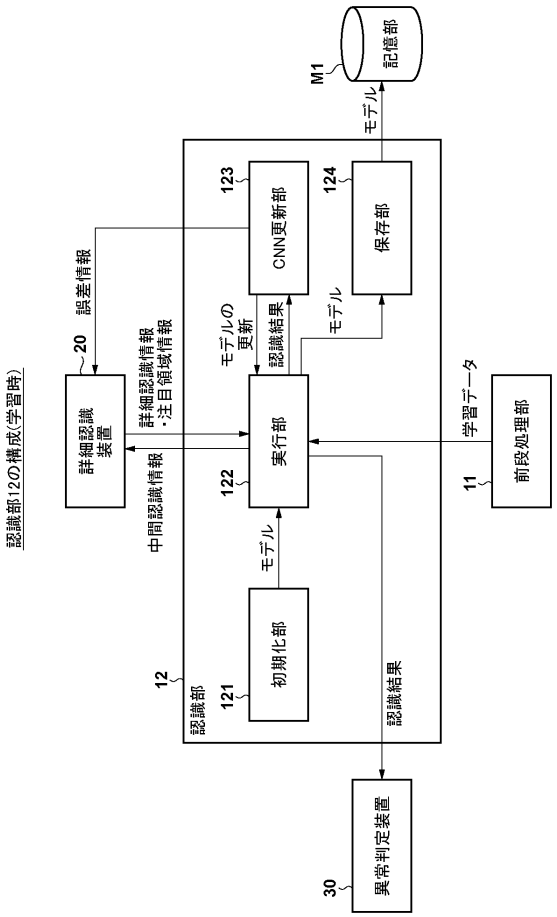
前段処理部11の動作(学習時)



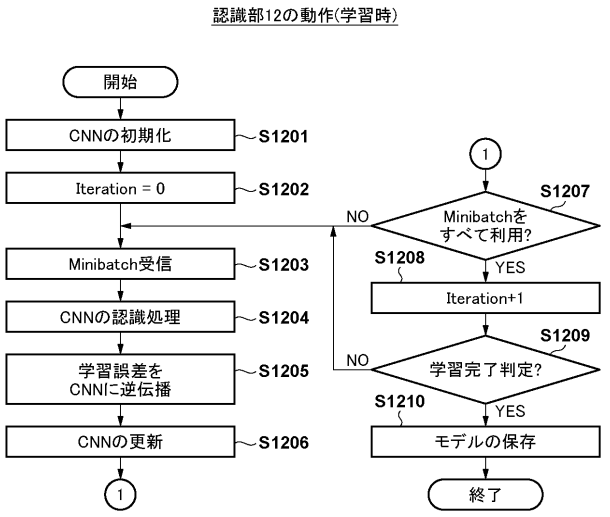
【図１０】



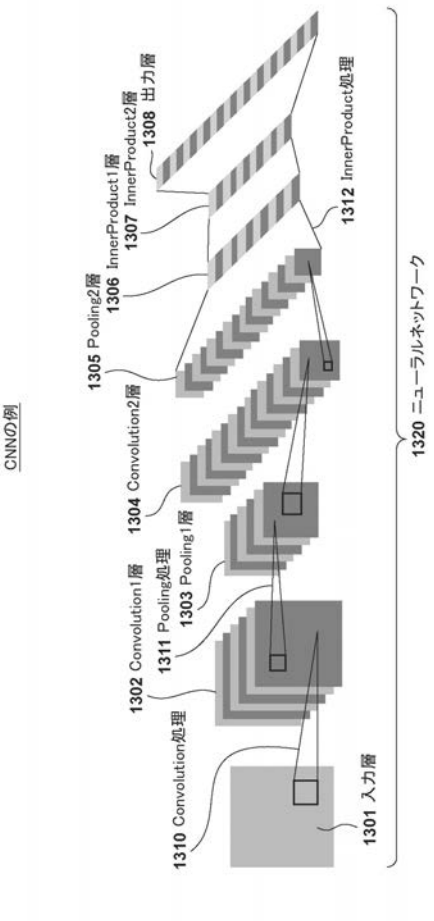
【図１１】



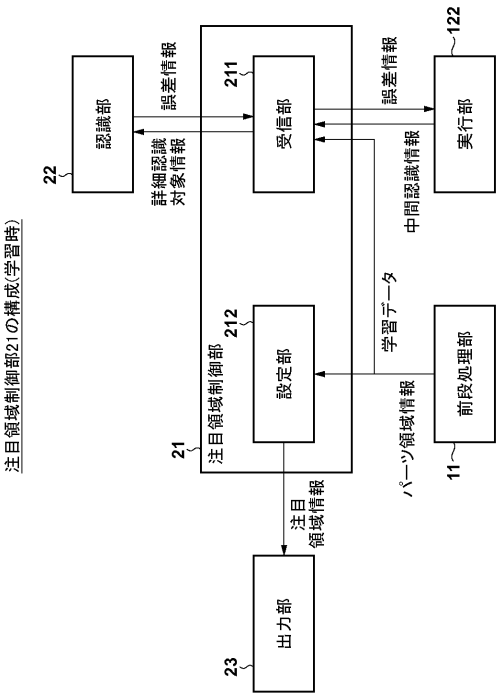
【図１２】



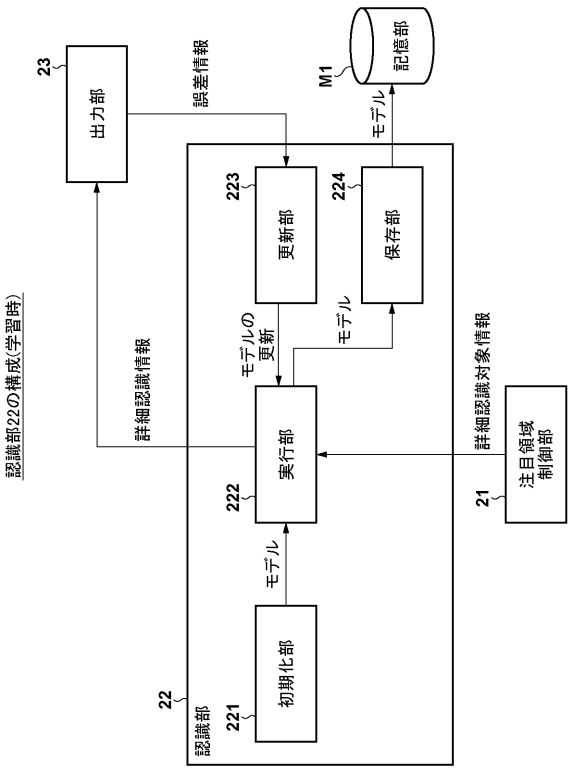
【図１３】



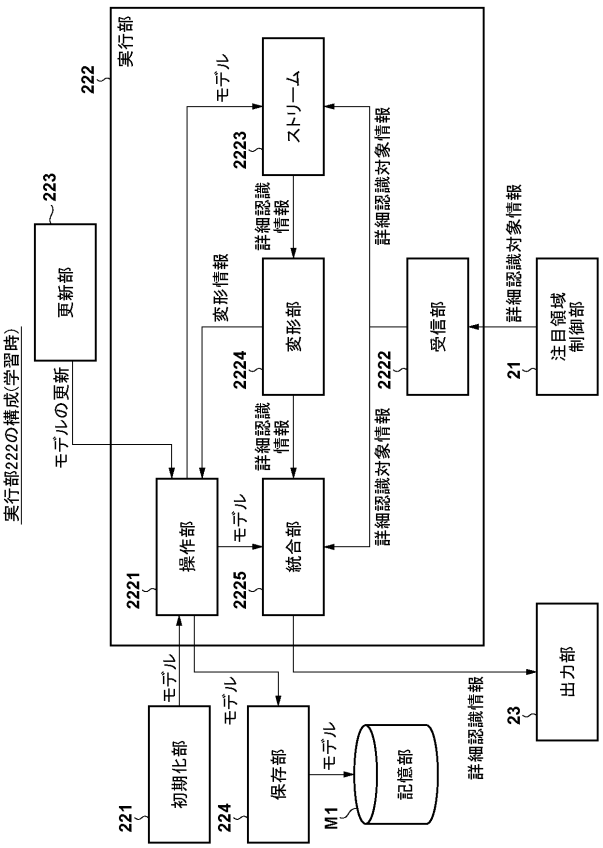
【図１４】



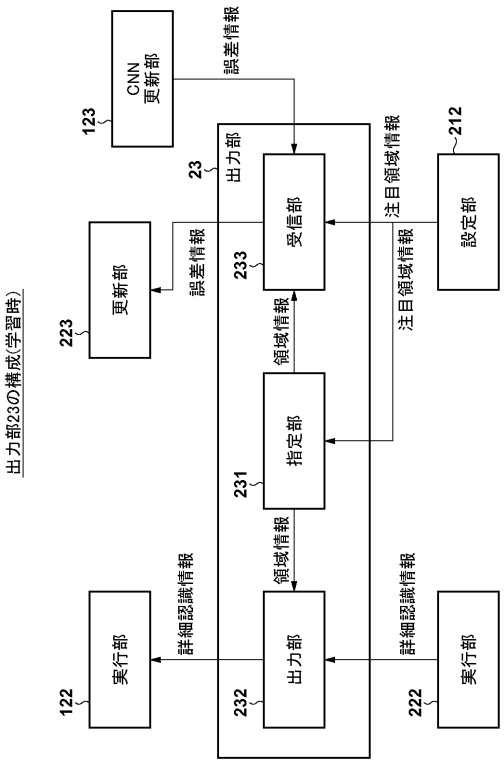
【図１５】



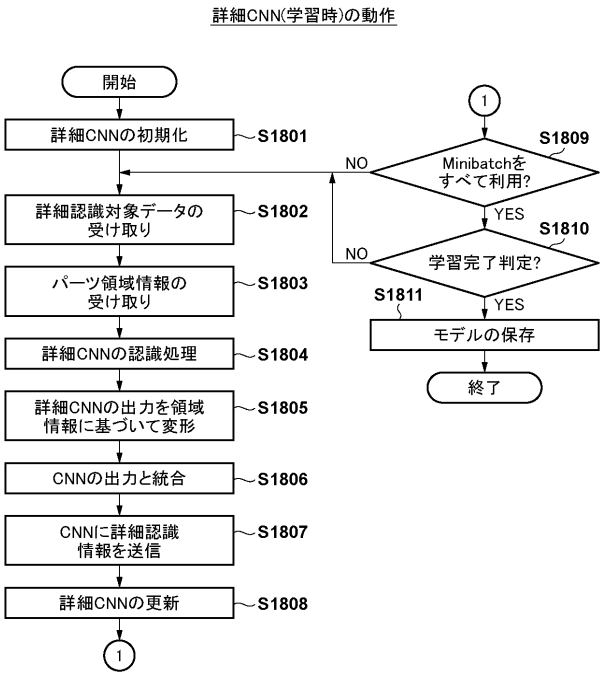
【図１６】



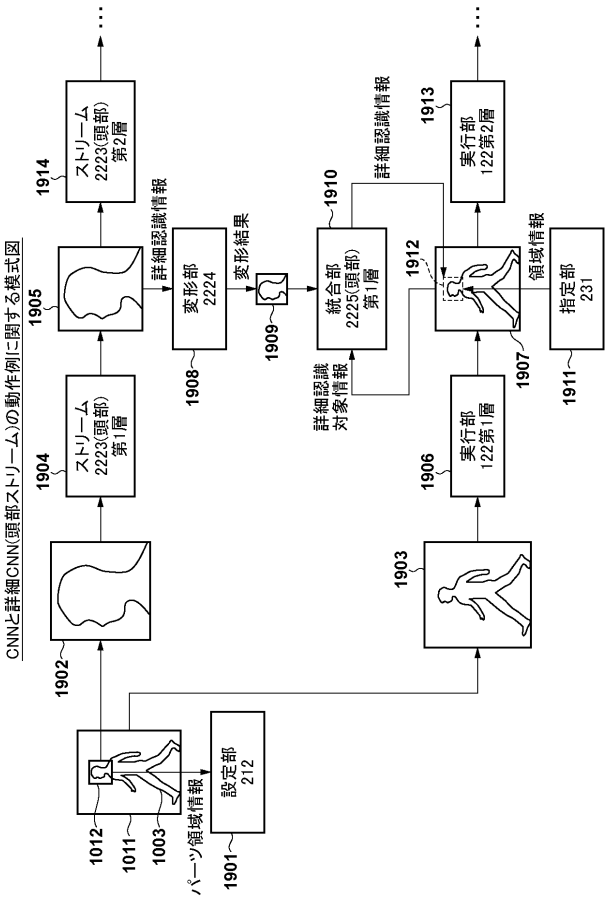
【図１７】



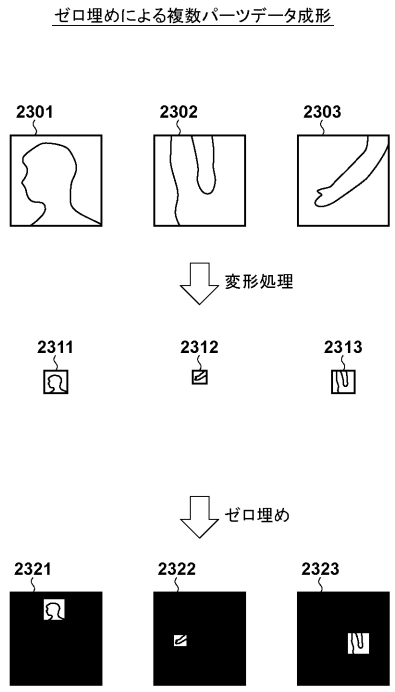
【図１８】



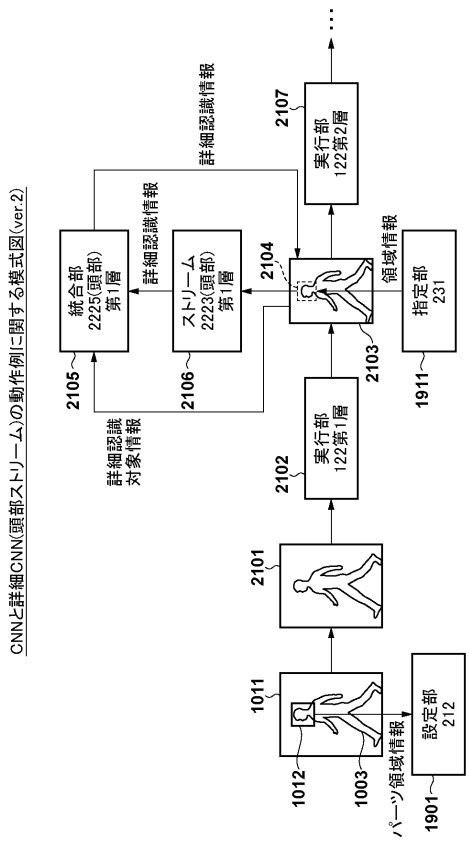
【図１９】



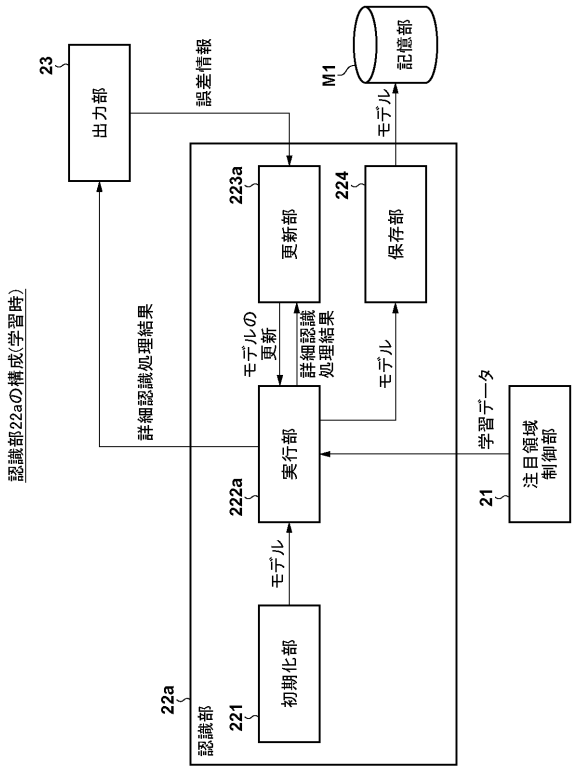
【図２０】



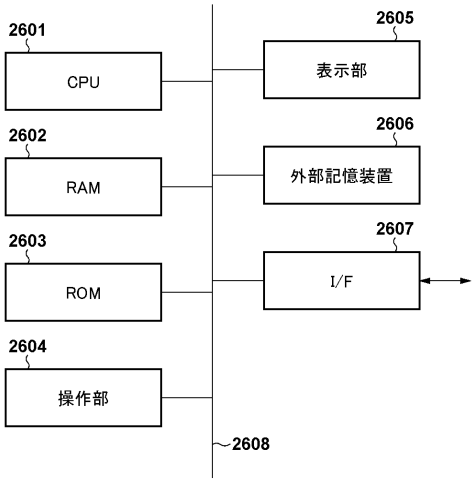
【図２１】



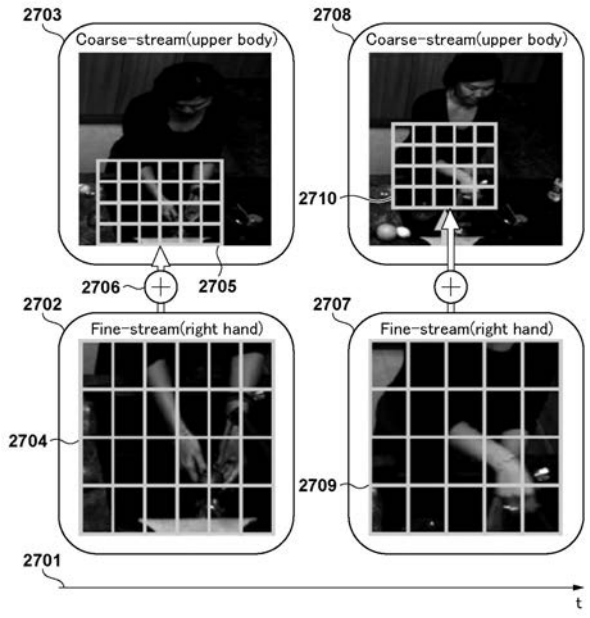
【図２２】



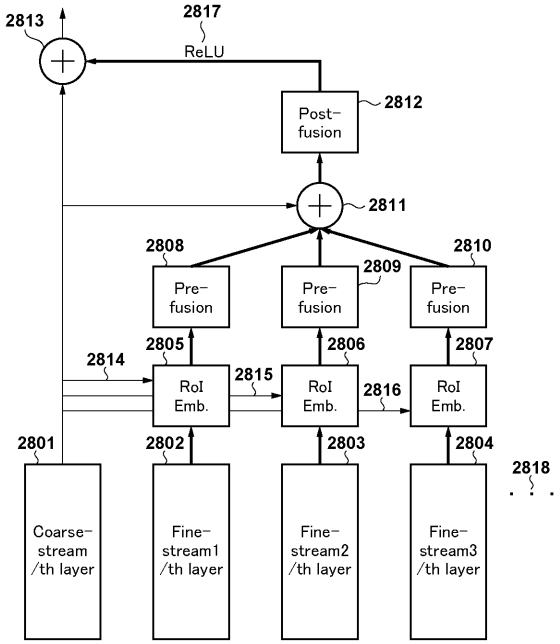
【図２３】



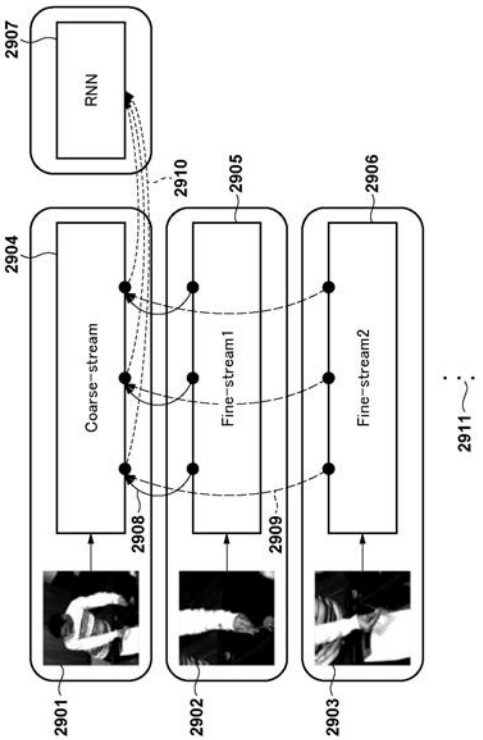
【図２４】



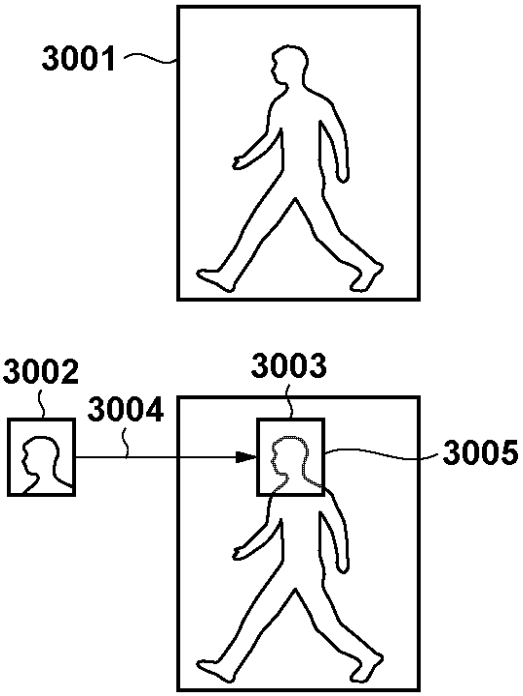
【図２５】



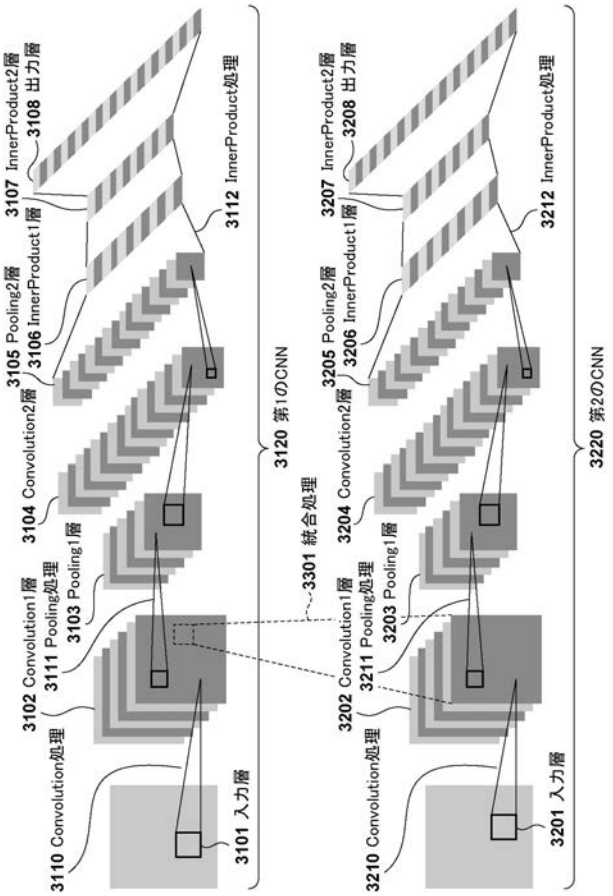
【図２６】


2907
RNN
2910
2905
2906
2904
Coarse-stream
Fine-stream1
Fine-stream2
2911
2908
2909
2901
2902
2903

【図２７】


3001
3002
3003
3004
3005

【図２８】


3107 InnerProduct2層
3108 出力層
3112 InnerProduct処理
3105 Pooling2層
3106 InnerProduct1層
3104 Convolution2層
3120 第1のCNN
3301 統合処理
3102 Convolution1層
3111 Pooling処理
3103 Pooling1層
3110 Convolution処理
3101 入力層
3207 InnerProduct2層
3208 出力層
3212 InnerProduct処理
3205 Pooling2層
3206 InnerProduct1層
3204 Convolution2層
3220 第2のCNN
3202 Convolution1層
3211 Pooling処理
3203 Pooling1層
3210 Convolution処理
3201 入力層

フロントページの続き

(72)発明者　斎藤　侑輝
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
(72)発明者　小森　康弘
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

Ｆターム(参考) 5L096 BA02 BA18 FA16 FA69 GA51 HA11 JA11 KA04 KA15