(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2018-190332
(P2018-190332A)
(43)公開日　平成30年11月29日(2018.11.29)

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

審査請求 未請求　請求項の数 18　ＯＬ　（全 20 頁）

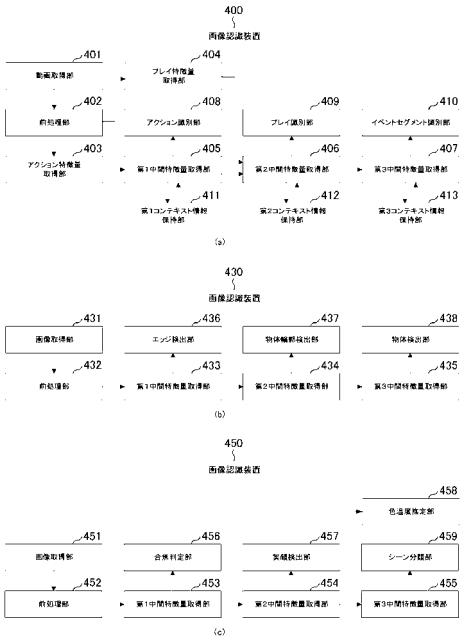
(21)出願番号　特願2017-94694(P2017-94694)
(22)出願日　平成29年5月11日(2017.5.11)

(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　100090273
弁理士　國分　孝悦
(72)発明者　角田　敬正
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　真継　優和
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 AA06　HA09　HA11　JA11　KA04

(54)【発明の名称】画像認識装置および学習装置

(57)【要約】
【課題】ＤＮＮにおいて、スケールが互いに異なる複数の認識タスクを有し、１つの中間層から識別層を分枝させたネットワークで認識処理を行うことができるようにする。
【解決手段】ニューラルネットワークを用いたネットワーク構造により対象の認識を行う画像認識装置であって、入力画像から第１の識別を行うための第１の中間特徴量を取得する第１の取得手段と、前記第１の中間特徴量に基づいて前記第１の識別を行う第１の識別手段と、前記第１の中間特徴量から前記第１の識別よりもスケールの大きい第２の識別を行うための第２の中間特徴量を取得する第２の取得手段と、前記第２の中間特徴量に基づいて前記第２の識別を行う第２の識別手段と、を有する。
【選択図】図４



【特許請求の範囲】
【請求項１】
ニューラルネットワークを用いたネットワーク構造により対象の認識を行う画像認識装置であって、
入力画像から第１の識別を行うための第１の中間特徴量を取得する第１の取得手段と、
前記第１の中間特徴量に基づいて前記第１の識別を行う第１の識別手段と、
前記第１の中間特徴量から前記第１の識別よりもスケールの大きい第２の識別を行うための第２の中間特徴量を取得する第２の取得手段と、
前記第２の中間特徴量に基づいて前記第２の識別を行う第２の識別手段と、
を有することを特徴とする画像認識装置。
【請求項２】
前記第１の識別手段は、前記ニューラルネットワークの中間層から分枝したネットワークによって実現されることを特徴とする請求項１に記載の画像認識装置。
【請求項３】
前記スケールは、時間的尺度、空間的尺度、あるいは問題の複雑さの尺度の何れかであることを特徴とする請求項１又は２に記載の画像認識装置。
【請求項４】
前記スケールは時間的尺度であり、前記ニューラルネットワークは再帰的ニューラルネットワーク（ＲＮＮ）であることを特徴とする請求項３に記載の画像認識装置。
【請求項５】
前記スケールは空間的尺度であり、前記ニューラルネットワークは畳み込みニューラルネットワーク（ＣＮＮ）であることを特徴とする請求項３に記載の画像認識装置。
【請求項６】
前記第２の識別手段は、さらに前記第１の中間特徴量に基づいて前記第２の識別を行うことを特徴とする請求項５に記載の画像認識装置。
【請求項７】
前記入力画像は、静止画又は動画であることを特徴とする請求項１～６の何れか１項に記載の画像認識装置。
【請求項８】
請求項１～７の何れか１項に記載の画像認識装置の識別器を学習する学習装置であって、
前記第１の取得手段および前記第２の取得手段をそれぞれ構成する中間層と前記第１の識別手段および前記第２の識別手段をそれぞれ構成する識別層とを初期化する初期化手段と、
前記初期化手段によって初期化された中間層および識別層を、学習データを用い学習する学習手段と、
前記学習手段による学習の結果に応じて、中間層もしくは識別層を新たに追加する追加手段と、
を有することを特徴とする学習装置。
【請求項９】
前記初期化手段は、前記中間層および前記識別層の数を最小限に構成したネットワーク構造を設定することを特徴とする請求項８に記載の学習装置。
【請求項１０】
前記追加手段は、前記学習手段による学習の結果に応じて、前記中間層の後段に新たな中間層を追加、もしくは前記中間層の最終層に新たな識別層を追加することを特徴とする請求項８又は９に記載の学習装置。
【請求項１１】
前記追加手段は、新たに中間層を追加する場合に前記識別層の下方に追加することを特徴とする請求項１０に記載の学習装置。
【請求項１２】

前記学習手段による学習の結果を、検証用データを用いて評価する評価手段をさらに有し、
前記評価手段による評価の結果、精度が閾値未満である場合に、前記追加手段は、新たに中間層を追加することを特徴とする請求項８～１１の何れか１項に記載の学習装置。
【請求項１３】
前記学習手段は、前記第１の識別手段および前記第２の識別手段による識別のそれぞれ対応する正解を含む学習データを用いて学習することを特徴とする請求項８～１２の何れか１項に記載の学習装置。
【請求項１４】
前記学習手段は、確率的勾配降下法を用いて学習を行うことを特徴とする請求項８～１３の何れか１項に記載の学習装置。
【請求項１５】
ニューラルネットワークを用いたネットワーク構造により対象の認識を行う画像認識装置の制御方法であって、
入力画像から第１の識別を行うための第１の中間特徴量を取得する第１の取得工程と、
前記第１の中間特徴量に基づいて前記第１の識別を行う第１の識別工程と、
前記第１の中間特徴量から前記第１の識別よりもスケールの大きい第２の識別を行うための第２の中間特徴量を取得する第２の取得工程と、
前記第２の中間特徴量に基づいて前記第２の識別を行う第２の識別工程と、
を有することを特徴とする画像認識装置の制御方法。
【請求項１６】
請求項１～７の何れか１項に記載の画像認識装置の識別器を学習する学習装置の制御方法であって、
前記第１の取得手段および前記第２の取得手段をそれぞれ構成する中間層と前記第１の識別手段および前記第２の識別手段をそれぞれ構成する識別層とを初期化する初期化工程と、
前記初期化工程において初期化された中間層および識別層を、学習データを用い学習する学習工程と、
前記学習工程による学習の結果に応じて、中間層もしくは識別層を新たに追加する追加工程と、
を有することを特徴とする学習装置の制御方法。
【請求項１７】
請求項１～７の何れか１項に記載の画像認識装置の各手段としてコンピュータを機能させるためのプログラム。
【請求項１８】
請求項８～１４の何れか１項に記載の学習装置の各手段としてコンピュータを機能させるためのプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、特に、スケールの異なる複数の認識タスクを実行するために用いて好適な画像認識装置、学習装置、画像認識装置の制御方法、学習装置の制御方法およびプログラムに関する。
【背景技術】
【０００２】
近年、画像や動画から物体などの対象を認識する技術が多様に研究されている。例えば、画像中の線分を抽出するエッジ検出や、類似する色やテクスチャの領域を切り出す領域分割、人物の目や口等のパーツを検出するパーツ検出、顔や人体、物体などを検出する物体認識、画像中の環境や状況を認識するシーン認識がある。ここで、認識処理の目的を認識タスクと呼ぶ。これらの認識対象は多様な尺度を持ち、個別の認識タスクを実行する認

識装置は、それぞれのスケールで定義された対象を、適切な特徴量に基づいて識別する。
【０００３】
近年、層の数が多くて深い多層ニューラルネットワーク（ＤＮＮ：Deep Neural Network）を用いた画像認識の研究が活発である。例えば非特許文献１には、畳み込みニューラルネットワーク（ＣＮＮ：Convolutional Neural Network）に正則化やデータ水増し等の様々な工夫を加えた技術が開示されている。この技術は、従来の課題とされていた過学習等を克服し、物体検出タスクにおいて非常に高い性能を出しており、注目されている。
【０００４】
また、非特許文献２には、ＣＮＮに関する詳細な調査が開示されている。このようなＣＮＮでは、入力層に近い低次の中間層で線分や単純図形、テクスチャ等の小さいスケールのプリミティブな特徴が捉えられている。また、識別層に近い高次の中間層で物体の形状やより詳細な特徴等、より大きいスケールの認識が行われる。
【０００５】
ＣＮＮでは主に画像をターゲットとするが、一方で、再帰的ニューラルネットワーク（ＲＮＮ：Recurrent Neural Networks）を用い、音声やテキスト等の時系列情報を認識する研究も盛んである。非特許文献３には、音声から音素の系列を認識するタスクに数種類のＲＮＮを適用し、精度を比較評価している技術が開示されている。
【先行技術文献】
【非特許文献】
【０００６】
【非特許文献１】Krizhevsky,A.,Sutskever,I.,&Hinton,G.E.,"Imagenet classification with deep convolutional neural networks.",In Advances in neural information processing systems(pp.1097-1105),2012.
【非特許文献２】Bolei Zhou,Aditya Khosla,Agata Lapedriza,Aude Oliva,Antonio Torralba,"Object detectors emerge in deep scene CNNs",ICLR2015
【非特許文献３】Alex Graves,Abdel-rahman Mohamed and Geoffrey Hinston,"Speech recognition with deep recurrent reural networks",Computing Research Repository 2013
【非特許文献４】J Donahue,Y Jia,O Vinyals,J Hoffman,N Zhang,E Tzeng,T Darrell,T Darrell,"DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition",arXiv2013
【非特許文献５】Xi Li, Liming Zhao, Lina Wei, MingHsuan Yang, Fei Wu, Yueting Zhuang, Haibin Ling, Jingdong Wang,"DeepSaliency: Multi-Task Deep Neural Network Model for Salient Object Detection",IEEE Transactions on Image Processing,2015
【非特許文献６】Karen Simonyan,Andrew Zisserman,"Two-Stream Convolutional Networks for Action Recognition in Videos",NIPS2015
【非特許文献７】Khurram Soomro, Amir Roshan Zamir, Mubarak Shah,"UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild", CRCV-TR-12-01,2012
【非特許文献８】Li Shen,Teck Wee Chua,Karianto Leman,"Shadow optimization from structured deep edge detection",CVPR2015
【発明の概要】
【発明が解決しようとする課題】
【０００７】
以上の研究により、認識タスクが定義するカテゴリの空間的および時間的スケールなどの様々な尺度や粒度を考えた場合、小さい粒度の認識タスクでは低階層のニューラルネットワーク（ＮＮ）で十分である。一方、大きい粒度の認識タスクでは高階層のＮＮが望ましい。
【０００８】
一方、ＣＮＮの中間層の特徴量を他の認識タスクに転用する研究も多く行われ（例えば非特許文献４参照）、認識タスクに依存する適切な層の存在は知られている。また、識別

層を分枝させ複数の認識タスクを１つのネットワークで実現する研究（例えば非特許文献５参照）も行われてきた。

【０００９】

しかし、前述したように、認識タスクによって時間的や空間的などのスケールは様々であり、さらには場合によっては、中間層の特徴量を他の認識タスクに転用し、１つの中間層から識別層を分枝させることが好ましい場合もある。このような場合に１つのネットワーク構造を実現するのが困難であった。

【００１０】

本発明は前述の問題点に鑑み、ＤＮＮにおいて、スケールが互いに異なる複数の認識タスクを有し、１つの中間層から識別層を分枝させたネットワークで認識処理を行うことができるようにすることを目的としている。

【課題を解決するための手段】

【００１１】

本発明に係る画像認識装置は、ニューラルネットワークを用いたネットワーク構造により対象の認識を行う画像認識装置であって、入力画像から第１の識別を行うための第１の中間特徴量を取得する第１の取得手段と、前記第１の中間特徴量に基づいて前記第１の識別を行う第１の識別手段と、前記第１の中間特徴量から前記第１の識別よりもスケールの大きい第２の識別を行うための第２の中間特徴量を取得する第２の取得手段と、前記第２の中間特徴量に基づいて前記第２の識別を行う第２の識別手段と、を有することを特徴とする。

【発明の効果】

【００１２】

本発明によれば、ＤＮＮにおいて、スケールが互いに異なる複数の認識タスクを有し、１つの中間層から識別層を分枝させたネットワークで認識処理を行うことができる。

【図面の簡単な説明】

【００１３】

【図１】動画の１フレームおよびカメラ配置の例を示す図である。

【図２】時間的スケールが互いに異なる各ラベルの関係を説明するための図である。

【図３】各実施形態に係る画像認識装置のハードウェア構成例を示すブロック図である。

【図４】各実施形態に係る画像認識装置の機能構成例を示すブロック図である。

【図５】各実施形態における画像認識装置による処理手順の一例を示すフローチャートである。

【図６】第１の実施形態に係るネットワーク構造の例を示す図である。

【図７】各実施形態に係る学習装置の機能構成例を示すブロック図である。

【図８】各実施形態における学習装置による処理手順の一例を示すフローチャートである。

【図９】最も浅いネットワークに構成した具体的な初期構造の例を示す図である。

【図１０】異なる識別層からの由来に基づく勾配を説明するための図である。

【図１１】認識タスクの精度が不十分でる場合の対応を説明するための図である。

【図１２】識別層を追加する場合の処理を説明するための図である。

【図１３】ＣＮＮのネットワーク構造の例を示す図である。

【図１４】ＣＮＮにおける特徴抽出ユニットと識別ユニットとを説明するための図である。

【図１５】第２の実施形態におけるＣＮＮのネットワーク構造の例を示す図である。

【図１６】第３の実施形態におけるＣＮＮのネットワーク構造の例を示す図である。

【発明を実施するための形態】

【００１４】

（第１の実施形態）

以下、本発明の第１の実施形態について、図面を参照しながら説明する。

例えば、サッカーやラグビーなどのチームスポーツ競技の未編集動画から、試合内容に

即したハイライトビデオを自動作成するには、競技特有のプレイや、それらの連鎖である重要なイベントの時間的セグメントを認識することが重要である。さらに競技特有のプレイを詳細に認識するには、映像全体から得られる特徴と、各プレイヤーのプリミティブな行動とが、重要な手がかりとなる。すなわち、試合内容に即したハイライトビデオの自動生成を実現するために、複数の認識タスクを実行することが重要である。つまり、プレイヤーの行動認識、競技特有のプレイ認識、重要イベントの時間的セグメントの認識（イベントセグメント認識）という、少なくとも３つの時間的尺度の異なる階層的な認識タスクを実行することが重要である。

【００１５】

各認識タスクは、上位のタスクほど時間的スケールの大きいタスクで、下位のタスクほど時間的スケールの小さいタスクであり、上位は下位を包含するため、１つのフレームワークで同時に実行することが効率的であるといえる。そこで、本実施形態では、例えばフットサル動画を題材に、ＲＮＮで認識時の処理として多層ＢＬＳＴＭ（Bidirectional Long Short-Term Memory）で異なる時間的尺度のマルチタスクを行う例について説明する。そして、学習時の処理として、多層ＢＬＳＴＭのネットワーク構造を最適化する方法について述べる。

【００１６】

図１（ａ）は、本実施形態において取得されるフットサル動画の１フレーム（静止画）１０１の例を示す図である。図１（ａ）において、動画の１フレーム１０１中には複数人のプレイヤー１０２とボール１０４とが存在する。また、図１（ｂ）に示すカメラ配置例１５１のように、コート１５２の周囲に複数台のカメラ１５３が配置され、全てのプレイヤーは何れかのカメラでキャプチャされるものとする。

【００１７】

図２は、動画（静止画の時系列）２０１に対する、時間的スケールが互いに異なる「イベントラベル」、「プレイラベル」、「アクションラベル」の関係を説明するための図である。アクションラベル２０５は、最も時間的スケールの小さいラベルで、スポーツ競技に依存せず汎用的なアクションを表現するラベルである。競技にはＮ人のプレイヤー２１１～２１Ｎが存在し、アクションラベル２０５はプレイヤー毎に設定される。プレイラベル２０４はアクションラベル２０５より時間的スケールが大きく、ボールを中心として定義される競技固有のプレイを表現するラベルである。イベントラベル２０３は最も時間的スケールの大きいラベルで、ゴールシーンやシュートシーンなどの重要シーンの時間的セグメントを表現するラベルである。以上のラベルは、画面やプレイヤーに対して常に１つのラベルが設定され、複数存在しないものとする。なお、以下の表１には、アクションラベル、プレイラベル、イベントラベルの例を示す。

【００１８】

【表１】

アクションラベル	プレイラベル	イベントラベル
キック	パス	シュート
レシーブ	シュート	ゴール
ドリブル	ゴール	コーナーキック
トラップ	ドリブル	フリーキック
キープ	クリア	
キャッチ	コーナーキック	
カット	フリーキック	
ラン	パントキック	
ターン	インターセプト	
ステップ	キーパーセーブ	
スライディング		

【００１９】
図３は、本実施形態に係る画像認識装置３００のハードウェア構成例を示すブロック図である。
図３において、画像認識装置３００は、ＣＰＵ３０１と、ＲＯＭ３０２と、ＲＡＭ３０３と、ＨＤＤ３０４と、表示部３０５と、入力部３０６と、ネットワークＩ／Ｆ部３０７とを有している。ＣＰＵ３０１は、ＲＯＭ３０２に記憶された制御プログラムを読み出して各種処理を実行する。ＲＡＭ３０３は、ＣＰＵ３０１の主メモリー、ワークエリア等の一時記憶領域として用いられる。ＨＤＤ３０４は、各種データや各種プログラム等を記憶する。表示部３０５は、各種情報を表示する。入力部３０６は、キーボードやマウスを有し、ユーザによる各種操作を受け付ける。
ネットワークＩ／Ｆ部３０７は、ネットワークを介して画像形成装置等の外部装置との通信処理を行う。また、他の例としては、ネットワークＩ／Ｆ部３０７は、無線により外部装置との通信を行ってもよい。
【００２０】
なお、後述する画像認識装置の機能や処理は、ＣＰＵ３０１がＲＯＭ３０２又はＨＤＤ３０４に格納されているプログラムを読み出し、このプログラムを実行することにより実現されるものである。また、他の例としては、ＣＰＵ３０１は、ＲＯＭ３０２等に替えて、ＳＤカード等の記録媒体に格納されているプログラムを読み出してもよい。
【００２１】
図４（ａ）は、本実施形態に係る画像認識装置４００の機能構成例を示すブロック図である。
本実施形態の画像認識装置４００は、動画取得部４０１、前処理部４０２、アクション特徴量取得部４０３、及びプレイ特徴量取得部４０４を有する。さらに、第１中間特徴量取得部４０５、第２中間特徴量取得部４０６、第３中間特徴量取得部４０７、アクション識別部４０８、プレイ識別部４０９、及びイベントセグメント識別部４１０を有する。さらに本実施形態における画像認識装置４００は、記憶手段として、第１コンテキスト情報保持部４１１、第２コンテキスト情報保持部４１２、第３コンテキスト情報保持部４１３を有する。画像認識装置４００が有するこれらの各機能の詳細については、図５（ａ）等を用いて後述する。

【００２２】
図５（ａ）は、本実施形態における画像認識装置４００による処理手順の一例を示すフローチャートである。
まず、Ｓ５０１において、動画取得部４０１は、複数のフレームからなる静止画像系列を取得する。例えば、図１（ａ）で例示した識別対象の動画の一部を動画取得部４０１が取得する。
【００２３】
本実施形態では、動画としてフットサルの試合映像を対象とし、最終的な動画の長さはフットサルの試合時間に準ずる。また、動画のフレームレートは、秒間３０フレーム程度の標準的な速度を想定する。本工程では、動画全体の中から、連続する数フレームから数１０フレームの予め設定した長さ（例えば試合時間）の静止画の系列を取得する。ここでは、説明を簡略化するため６０フレームの静止画を取得するものとする。この６０フレームは、外部装置に予め記憶されていた動画から取得しても良いが、リアルタイムにカメラで撮影した動画から取得してもよい。前者の場合、動画取得部４０１は、外部装置に記憶された試合映像から所定のフレーム数の静止画系列を取得し、後者の場合、カメラが出力する静止画系列から所定のフレーム数を取得する。また、図１（ｂ）に示したように、複数台のカメラで撮影されたことを前提とし、本工程ではそのカメラ台数分の静止画系列を取得する。
【００２４】
次に、Ｓ５０２においては、前処理部４０２は、Ｓ５０１で取得した静止画系列に対して前処理を行う。まず、前処理部４０２は、カメラ台数分の静止画系列に映る各プレイヤーとボールとをトラッキングし、プレイヤー位置およびプレイヤー領域、ボール位置の情報を取得する。ここで取得するプレイヤー領域は、図１（ａ）に示したプレイヤー領域１０３に該当する領域である。
【００２５】
次に、前処理部４０２は、画面全体と各プレイヤー領域とに関して、図１（ｂ）のように配置された複数台のカメラからよりよい配置のカメラを選択する。本実施形態では、画面全体に関しては最初のフレームでボールをより手前で撮影しているカメラ、プレイヤー領域に関しては最初のフレームで領域の隠れがなくプレイヤーがより大きく映っているカメラを選択する。次に、前処理部４０２は、画面全体および各プレイヤー領域の静止画系列の最初の１フレームに対し、色成分の正規化を行い、さらに静止画系列からオプティカルフローを取得するための処理を行う。オプティカルフローは、ここではフレーム間のブロックマッチング法を用いて取得する。
【００２６】
次に、Ｓ５０３においては、アクション特徴量取得部４０３は、プレイヤー毎のアクションを識別するためのアクション特徴量を取得する。ここでアクション特徴量とは、動画取得部４０１で取得した静止画系列から計算される、少なくともアピアランスとモーションとの２つに関する特徴量である。非特許文献６には、アピアランスおよびモーションに対応する２つのＣＮＮを用いて、数秒から数分程度の動画で構成される公開データセット（非特許文献７参照）、ＵＣＦ－１０１等に対するアクション分類が開示されている。
【００２７】
モーションに関するＣＮＮには、オプティカルフローのベクトル場を鉛直方向成分と垂直方向成分とに分けた２つのマップを１単位として、それらを積み上げた複数パターン（１、５、１０等）を入力としている。本実施形態では、１０単位分のオプティカルフローを入力とし、ＵＣＦ－１０１のアクション分類タスクで学習したＣＮＮを用い、Ｓｏｆｔｍａｘ識別層の前の全結合層の出力をモーションに関する特徴量として用いる。
【００２８】
一方、アピアランスに関する特徴量は、非特許文献４に記載された方法と同様に、物体検出タスクの公開データセットであるＩＬＳＶＲＣを用いて学習したＣＮＮのＳｏｆｔｍａｘ識別層の前の全結合層の出力を用いる。予め学習したＣＮＮの中間出力を別の認識タ

スクに転用するこのような利用方法は、非特許文献４等で記載の通り一般的な方法であるので、これ以上の詳細な説明は割愛する。

【００２９】

次に、Ｓ５０４においては、プレイ特徴量取得部４０４は、正規化した画面全体の静止画とオプティカルフローとから、プレイを識別するためのプレイ特徴量を取得する。ここでは、アクション特徴量取得部４０３が行った処理と同様の方法でアピアランスおよびモーションに関する特徴量をプレイ特徴量として取得する。

【００３０】

次に、画面全体から取得したプレイ特徴量と、プレイヤー毎のアクション特徴量とから、アクション、プレイ、イベントセグメントの３つを識別する。第１中間特徴量取得部４０５、第２中間特徴量取得部４０６、および第３中間特徴量取得部４０７は、後述する学習時の処理で最適化された層数を持つＢＬＳＴＭに対応する。

【００３１】

図６は、アクション、プレイ、イベントセグメントに関してＢＬＳＴＭ等のユニットでグラフ表現した模式図である。図６に示すように、プレイヤー毎のアクション特徴量６０１は、それぞれ独立に複数層で構成されるＢＬＳＴＭ６０５に入力され、アクション識別層６０６を介して、プレイヤー毎のアクションが識別される。そして画面全体から取得したプレイ特徴量６０４は、複数層のＢＬＳＴＭ６０５を介して、プレイヤー毎のアクション特徴量６０１と統合される。そして複数層のＢＬＳＴＭ６０５とプレイ識別層６０７とを介してプレイラベルが識別される。プレイ特徴量６０４およびアクション特徴量６０１は、さらに複数層のＢＬＳＴＭ６０５とイベントセグメント識別層６０８とを介し、最終層でイベントラベルが識別される。ここで、アクション識別層６０６およびプレイ識別層６０７は、ソフトマックス識別器と中間層とを１層ずつ持つマルチレイヤーの識別層であり、イベントセグメント識別層６０８はソフトマックス識別器で構成されている識別層である。

【００３２】

まず、Ｓ５０５において、第１中間特徴量取得部４０５は、アクション特徴量の時系列に関する特徴量を取得する。第１中間特徴量取得部４０５は、ＢＬＳＴＭによって構成され、第１コンテキスト情報保持部４１１には、前ステップの第１中間特徴量が保存されている。第１中間特徴量取得部４０５は、アクション特徴量取得部４０３で取得したアクション特徴量と前ステップの第１中間特徴量とから現ステップの第１中間特徴量を計算する。また、このＢＬＳＴＭには、後述する学習時の処理で記載した方法に従って学習したＢＬＳＴＭを用いる。

【００３３】

次に、Ｓ５０６において、アクション識別部４０８は、現ステップの第１中間特徴量に、ソフトマックス識別器を用いることで、アクションラベルの識別スコアを計算する。このソフトマックス識別器の学習方法に関しては後述する。

【００３４】

Ｓ５０７においては、第２中間特徴量取得部４０６は、プレイ特徴量の時系列に関する特徴量を取得する。第２中間特徴量取得部４０６は、ＢＬＳＴＭによって構成され、第２コンテキスト情報保持部４１２には、前ステップの第２中間特徴量が保存されている。第２中間特徴量取得部４０６は、プレイ特徴量取得部４０４で取得したプレイ特徴量と、現ステップの第１中間特徴量と、前ステップの第２中間特徴量とを演算して現ステップの第２中間特徴量を計算する。また、このＢＬＳＴＭには、後述する学習時の処理で記載した方法に従って学習したＢＬＳＴＭを用いる。

【００３５】

次に、Ｓ５０８において、プレイ識別部４０９は、現ステップの第２中間特徴量に、ソフトマックス識別器を用いることで、プレイラベルの識別スコアを計算する。

【００３６】

Ｓ５０９においては、第３中間特徴量取得部４０７は、現ステップの第３中間特徴量を

取得する。第３中間特徴量取得部４０７は、ＢＬＳＴＭによって構成され、第３コンテキスト情報保持部４１３には、前ステップの第３中間特徴量が保存されている。第３中間特徴量取得部４０７は、現ステップの第２中間特徴量と前ステップの第３中間特徴量とを演算して現ステップの第３中間特徴量を計算する。また、このＢＬＳＴＭには、後述する学習時の処理で記載した方法に従って学習したＢＬＳＴＭを用いる。

【００３７】

次に、Ｓ５１０において、イベントセグメント識別部４１０は、現ステップの第３中間特徴量に、ソフトマックス識別器を用いることで、イベントラベルの識別スコアを計算する。

【００３８】

次に、本実施形態におけるＳ５０５～Ｓ５１０で利用するＢＬＳＴＭ（中間層）および識別層の学習方法と、図６に示したネットワーク構造の最適化方法とについて説明する。

【００３９】

図７は、本実施形態における学習装置７００の機能構成例を示すブロック図である。なお、本実施形態に係る学習装置７００のハードウェア構成については、基本的には図３に示した構成と同様の構成である。

図７に示すように、学習装置７００は、ネットワーク構造初期化部７０１、ネットワークパラメータ最適化部７０２、精度評価部７０３、層追加部７０４、およびタスク追加部７０５を有する。さらに学習装置７００は、記憶手段として、学習用データ保持部７０６、検証用データ保持部７０７、およびネットワークパラメータ保持部７０８を有する。

【００４０】

図８は、本実施形態における学習装置７００が行う処理手順の一例を示すフローチャートである。ここで各工程の概要及び図７に示した各構成の機能について説明する。

Ｓ８０１においては、ネットワーク構造初期化部７０１は、認識タスク毎の認識層とネットワーク全体の構造とを初期化する。本実施形態では、まず認識タスクとして、最もスケールの小さいアクション認識と次にスケールの小さいプレイ認識とを、最も浅いネットワークに構成した構造を初期構造とする。図９には、最も浅いネットワークに構成した具体的な初期構造の例を示す。図６に示した構造と比較して、アクション特徴量９０１、プレイ特徴量９０４、ＢＬＳＴＭ９０５、アクション識別層９０６およびプレイ識別層９０７に示すように、中間層及び識別層の数を最小限にすることにより、最も浅いネットワークを構成している。

【００４１】

次に、Ｓ８０２において、ネットワークパラメータ最適化部７０２は、学習用データ保持部７０６に保持されている学習用データを用い、ネットワークを構成するＢＬＳＴＭおよび識別層のパラメータの最適化（学習）を行う。ここで、最適化するパラメータのうち、プレイヤー毎のアクション識別層およびそれに接続されるＢＬＳＴＭは、プレイヤーが変わっても同じ定義のラベルを識別し、プレイヤーの見えは試合毎に様々なバリエーションがあり得る。そこで、プレイヤー毎のＢＬＳＴＭおよびアクション識別層において重みは共有するものとする。

【００４２】

ここで、学習用データには、動画を構成する静止画の時系列に対応させて、イベントラベル、プレイラベル、プレイヤー毎のアクションラベルの正解ラベルが用意されている。各ＢＬＳＴＭおよび識別層のパラメータを最適化する際には確率的勾配降下法を用いる。この場合、図１０に示すように、ＢＬＳＴＭ１００１において、プレイ識別層１００３に由来する勾配１００５と、アクション識別層１００２に由来する勾配１００４とが混合する。このようなＢＬＳＴＭでは、複数の勾配を単純に平均しても良いが、検証用データを用意して誤差率を計量し、その誤差率で勾配を混合させても良い。

【００４３】

誤差逆伝搬法で計算される当該ＢＬＳＴＭの全体の勾配をΔＥとすると、アクション識別層に由来する勾配をΔＥ$^{(1)}$、プレイ識別層に由来する勾配をΔＥ$^{(2)}$とした場合に、Ｂ

ＬＳＴＭの全体の勾配は以下の式（１）により算出される。

$$\Delta E = \alpha \Delta E^{(1)} + \beta \Delta E^{(2)} \quad \cdots (1)$$

【００４４】

ここで、αはアクション識別の検証用データに関する誤差率に比例する値で、βはプレイ識別の検証用データに関する誤差率に比例する値である。α，βは足して１となるように正規化されている。

【００４５】

また、パラメータの最適化には確率的勾配降下法を適用するため、学習データはミニバッチに分割するが、検証用データも同様にミニバッチに分割し、ミニバッチ毎に誤差率を計算しても良い。こうすることで、ミニバッチ毎に混合比が変わるため、ミニバッチ毎に勾配を変える確率的勾配降下法と同様の効果、すなわちより良い局所解に降下する可能性が増すと考えられる。

【００４６】

また、アクションの識別はプレイヤー毎に行うため、正解ラベルはＮ人のプレイヤー分あり、以下の式（２）によりＮ人のプレイヤー分の勾配が得られる。

$$\Delta E^{(1)} = \Delta E_1^{(1)} + \Delta E_2^{(1)} + \cdots + \Delta E_N^{(1)} \quad \cdots (2)$$

ＢＬＳＴＭの全体の勾配を算出する際に、式（１）を用いたが、代わりに以下の式（３）を用いてもよい。この場合、アクション識別の勾配は、プレイ識別の勾配に対しＮ人分の重みが加わっているので、それを相殺するためにＮでアクション識別の勾配を割っても良い。

$$\Delta E = (\alpha / N) \Delta E^{(1)} + \beta \Delta E^{(2)} \quad \cdots (3)$$

以上の処理により学習されたパラメータは、ネットワークパラメータ保持部７０８に記憶される。

【００４７】

次にＳ８０３において、精度評価部７０３は、検証用データ保持部７０７に記憶されている検証用データを用い、最適化されたパラメータに基づく各識別タスクの精度を得る。すなわち、Ｓ８０１およびＳ８０２を経た現段階では、アクション識別およびプレイ識別の２つのタスクに関して、精度を計算する。本工程で用いる検証用データは、Ｓ８０２での勾配の混合比の決定に用いる検証用データと同じものでも良いし、別の検証用データを用意しても良い。

【００４８】

次に、Ｓ８０４において、精度評価部７０３は、Ｓ８０３で得た精度が閾値以上であるか否かを判定する。精度が閾値未満で精度が不十分な場合はＳ８０５に分岐し、そうでない場合はＳ８０６に分岐する。

【００４９】

Ｓ８０５においては、層追加部７０４は、タスク毎に既に存在するＢＬＳＴＭの後段にＢＬＳＴＭを追加する。アクション識別、プレイ識別ともに精度が十分でない場合には図１１（ａ）に示すように、アクション識別層１１０６の下方と、プレイ識別層１１０５の下方とにＢＬＳＴＭ１１０１、１１０２を追加する。また、アクション識別の精度のみが十分でない場合は、図１１（ｂ）に示すように、アクション識別層の下方のみにＢＬＳＴＭ１１０３を追加する。一方、プレイ識別の精度のみが十分でない場合には、図１１（ｃ）に示すように、プレイ識別層の下方のみにＢＬＳＴＭ１１０４を追加する。そして、Ｓ８０２に戻る。

【００５０】

一方、識別タスクの精度がともに十分である場合、その識別タスクの識別層の分枝位置が決定する。そして、Ｓ８０６において、精度評価部７０３は、全てのタスクの精度が十分であるか否かを判定する。この判定の結果、精度が不十分であるタスクがある場合は、Ｓ８０７に分岐する。

【００５１】

Ｓ８０７においては、タスク追加部７０５は、タスクを追加する。そして、Ｓ８０５を

経て、追加するタスクのＢＬＳＴＭと識別層とを追加することになる。図１１（ａ）が現在の状態である場合、図１２に示すように、Ｓ８０７では最終層にイベントセグメント識別層１２０１が追加され、イベントセグメント識別層１２０１の下方にはＳ８０５でＢＬＳＴＭ１２０２が追加される。

【００５２】

以降、Ｓ８０２およびＳ８０３を再度行い、最終的に全てのタスクで精度が十分になるまで、ＢＬＳＴＭの追加と識別層の分枝位置の探索とを繰り返す。

【００５３】

以上のように本実施形態によれば、アクション識別、プレイ識別、イベントセグメント識別という時間的スケールの異なるマルチタスクを行うＲＮＮのニューラルネットワークを実現できる。これにより、スポーツ動画の入力に対してアクション、プレイ、イベントセグメントの各ラベルを同時に推定することができる。

【００５４】

（第２の実施形態）

本実施形態では、空間的尺度の異なるマルチタスクを行うＣＮＮを例に説明する。非特許文献２に記載の技術では、１０００クラスの物体検出タスクで学習した７層のＣＮＮにおいて各層が何に反応しているかをクラウドソーシングで調査している。受容野に対応する画像をワーカーに分類させると、１層目が色や単純図形、２層目がテクスチャ、３層目が領域や表面、４層目がオブジェクトパーツ、５層目がオブジェクトに反応しているとする答えが多かった。このことは最終的に１０００クラスの物体検出を実現するために、低層で暗黙的に上記の性質が獲得されていることを示している。最終層で物体検出を学習すると同時に、低次の層で単純図形やテクスチャ、領域分割等のプリミティブな画像認識を教師あり学習の枠組みで積極的に学習することで、１つのネットワークで空間的スケールの異なるマルチタスクを行うＣＮＮが学習できる。

【００５５】

そこで本実施形態では、コンピュータビジョンにとって特に重要で、かつ教師あり学習の仕組みで学習しやすい複数のタスクに関し、ニューラルネットワークの構造を決定して学習し、識別する方法を説明する。具体的には、「エッジ検出」、「物体輪郭検出」、「物体検出」の３つのタスクに関し、これを行うニューラルネットワークの構造を決定して学習し、識別する方法を説明する。

【００５６】

まず、本実施形態において学習されるＣＮＮを用いて画像を識別する際の処理について説明する。ＣＮＮは畳みこみ演算を多く行うニューラルネットワークであり、非特許文献１や非特許文献２に開示されている。つまり、ＣＮＮでは、畳み込み層（Ｃｏｎｖ（Convolutional Layer））と非線形処理（ＲｅＬＵ（Rectified linear unit）やＰｏｏｌ（Max pooling）など）との組み合わせで特徴抽出層が実現される。そのあと、ＣＮＮは全結合層（ＦＣ（Fully-connected Layer））を経て画像分類結果（各クラスに対する尤度）を出力する。

【００５７】

図１３に、非特許文献１に開示されているネットワーク構造の例を示す。画像Ｉｍｇ１３０１を入力すると、Ｃｏｎｖ＋Ｐｏｏｌ１３０２、Ｃｏｎｖ＋Ｐｏｏｌ１３０３、Ｃｏｎｖ１３０４、Ｃｏｎｖ１３０５、Ｃｏｎｖ＋Ｐｏｏｌ１３０６を順番に適用して特徴抽出が行われる。その後、識別層にてＦＣ１３０７、ＦＣ１３０８、Ｓｏｆｔｍａｘ１３０９を行い、マルチクラスのカテゴリ尤度を出力している。なお、全てのＣｏｎｖおよびＦＣの後には非線形処理（ＲｅＬＵ）を行うが、図１３に示す例では簡略化のために省略している。また、画像Ｉｍｇ１３０１は、ＣＮＮに入力する際、所定サイズに画像をリサイズしてクロップし、正規化するなどの前処理が行われるのが一般的である。

【００５８】

図４（ｂ）は、本実施形態の画像認識装置４３０の機能構成例を示すブロック図である。なお、本実施形態に係る画像認識装置４３０のハードウェア構成については図３と同様

である。
本実施形態の画像認識装置４３０は、画像取得部４３１、前処理部４３２、第１中間特徴量取得部４３３、第２中間特徴量取得部４３４、第３中間特徴量取得部４３５、エッジ検出部４３６、物体輪郭検出部４３７、および物体検出部４３８を有する。画像認識装置４３０が有するこれらの各機能の詳細については、図５（ｂ）等を用いて後述する。
【００５９】
図５（ｂ）は、本実施形態における画像認識装置４３０による処理手順の一例を示すフローチャートである。
まず、Ｓ５３１において、画像取得部４３１は、マルチタスクの認識を行う対象である静止画を取得する。静止画は外部の記憶装置から取得しも良いし、スチルカメラ等の撮像デバイスで新たに取得しても良い。
【００６０】
次にＳ５３２においては、前処理部４３２は、前工程で取得した入力画像に対し前処理を行う。ここでは入力画像に対して色成分を正規化して２５６×２５６程度の大きさにリサイズし、続いて２２４×２２４程度の大きさにクロップする前処理を行う。
【００６１】
次に前処理した入力画像から、第１中間特徴量取得部４３３、第２中間特徴量取得部４３４、第３中間特徴量取得部４３５の３つの中間特徴量取得部がそれぞれ特徴量を取得する。そして、エッジ検出部４３６、物体輪郭検出部４３７、および物体検出部４３８が、それぞれエッジ抽出、物体輪郭抽出、物体検出の３つを行う。３つの各中間特徴量取得部は、後述する学習時の処理で最適化された構造を持つＣＮＮである。
【００６２】
ここで、スケールの異なるマルチタスクを行うＣＮＮを実現するために、図１４に示したような基本ユニットを考える。図１４（ａ）～図１４（ｃ）は、それぞれＣｏｎｖとＰｏｏｌとが組み合わされた特徴抽出層のユニット（特徴抽出ユニット）の例を示している。より低層側では、図１４（ａ）に示す特徴抽出ユニットを用い、中層側では図１４（ｂ）に示す特徴抽出ユニットを用い、高層側では、図１４（ｃ）に示す特徴抽出ユニットを用いるものとする。
【００６３】
また、エッジ抽出および物体輪郭抽出では、ピクセル毎の２値分類を行うので、識別層として図１４（ｄ）に示すようなＦＣとLogistic Regressionとを組み合わせた識別ユニットを用いる。ＦＣは通常のように特徴抽出ユニット最終層の全チャネル全レスポンスマップと接続せずに、マップ上の４×４の領域をチャネル方向に全結合しマップ上の各点で２値判定を行う。そして、最終アウトプットである物体検出では、マルチクラスの分類を行うので、図１４（ｅ）に示すようなＦＣ×２＋Ｓｏｆｔｍａｘとなる識別ユニットを用いるものとする。
【００６４】
図１５は、図４（ｂ）の各構成および入力画像のネットワークを、図１４のユニットを用いて表現した図である。第１中間特徴量取得部４３３は図１４（ａ）のＣｏｎｖ×２＋Ｐｏｏｌの特徴抽出ユニットで実現される。そして、第２中間特徴量取得部４３４は図１４（ｂ）のＣｏｎｖ×３＋Ｐｏｏｌの特徴抽出ユニット、第３中間特徴量取得部４３５は図１４（ｃ）のＣｏｎｖ×４＋Ｐｏｏｌの特徴抽出ユニットで実現される。また、前述の通り、エッジ検出部４３６および物体輪郭検出部４３７は、それぞれ図１４（ｄ）のLogistic Regressionの識別ユニットで実現される。そして、物体検出部４３８は、図１４（ｅ）のＳｏｆｔｍａｘの識別ユニットで実現される。
【００６５】
図１５に示すように、入力画像１５０１から、特徴抽出ユニット１５０２を介して、Logistic Regressionの識別ユニット１５０５でエッジ検出が行われる。そして、特徴抽出ユニット１５０２のもう一方の出力により、特徴抽出ユニット１５０３を介し、Logistic Regressionの識別ユニット１５０６で物体輪郭抽出が行われる。さらに、特徴抽出ユニ

ット１５０３の出力により、特徴抽出ユニット１５０４を介し、Ｓｏｆｔｍａｘの識別ユニット１５０７で物体検出が行われる。

【００６６】

Ｓ５３３においては、第１中間特徴量取得部４３３は、前処理が行われた入力画像からエッジ検出に係る第１中間特徴量を取得する。そして、Ｓ５３４において、エッジ検出部４３６は、第１中間特徴量から、画像中のエッジを検出する。

【００６７】

次に、Ｓ５３５において、第２中間特徴量取得部４３４は、第１中間特徴量に基づき、物体の輪郭に係る第２中間特徴量を取得する。そして、Ｓ５３６において、物体輪郭検出部４３７は、第２中間特徴量から、物体の輪郭を検出する。

【００６８】

次に、Ｓ５３７において、第３中間特徴量取得部４３５は、第２の中間特徴量に基づき、物体検出に係る第３中間特徴量を取得する。そして、Ｓ５３８において、物体検出部４３８は、第３中間特徴量から、画像中の物体を検出する。

【００６９】

以上のように、Ｓ５３３、Ｓ５３５、Ｓ５３７およびＳ５３８における処理は、一般的なＣＮＮにおける認識時の処理と同様であるので、詳細な説明は割愛する。そして、Ｓ５３４およびＳ５３６の処理は、画素毎の２値分類を行う処理であるが、ここでは非特許文献８に記載の方法を参考にして実現する。すなわち、特徴抽出ユニットの最終層のレスポンスマップ上の４×４の領域を、チャネル方向に２５次元のＦＣと全結合させる。そして、レスポンスマップ上の４×４領域に対応する受容野において、中央に位置する５×５ピクセルの全点でエッジの有無を２値判定するという方法である。

【００７０】

以上の処理により、入力画像に対し、エッジ抽出、物体輪郭抽出、物体検出の各識別タスクを１つのネットワークで同時に実現できる。次に、本実施形態におけるＳ５３３～Ｓ５３８で利用する特徴抽出ユニットおよび識別ユニットの学習方法と図１５に示したネットワーク構造の探索方法とについて説明する。なお、本実施形態における学習装置の構成は、第１の実施形態と同様であり、本実施形態における学習の処理手順は、基本的には図８と同様である。以下、第１の実施形態との相違について説明する。

【００７１】

まず、Ｓ８０１において、ネットワーク構造初期化部７０１は、ＣＮＮのネットワーク構造を初期化する。次のＳ８０２においては、ネットワークパラメータ最適化部７０２は、エッジ検出、物体輪郭検出（および物体検出）のネットワークに組み込まれたタスクの識別ユニットと特徴抽出ユニットとのネットワークパラメータを最適化する。Ｓ８０３では、精度評価部７０３は、エッジ検出、物体輪郭検出（および物体検出）の各タスクの評価用データを用い、タスク毎に精度を評価する。Ｓ８０５では、層追加部７０４は、精度が不十分であるタスクに応じて、図１４（ａ）～図１４（ｃ）で例示した特徴抽出ユニットを追加する。Ｓ８０７では、タスク追加部７０５は、図１４（ｅ）に示す物体検出用の識別ユニットを追加する。

【００７２】

ここで、学習用データは、第１の実施形態と同様に、エッジ検出、物体輪郭検出、物体検出の３つのタスクの正解ラベルが各学習用画像に付けられた物を用いても良い。一方、エッジ検出、物体輪郭検出、物体検出のタスク毎に別々に学習用画像と正解ラベルとを用意しても良い。前者の場合、Ｓ８０２では、第１の実施形態と同様に勾配を混合させ、パラメータの最適化を実行すれば良い。後者の場合、Ｓ８０２において、非特許文献５に開示された方法と同様にタスク毎にロス関数を切り替え、最適化すれば良い。

【００７３】

また、後者の場合、同じタスクの公開データセットを利用しても良い。例えば物体検出ではLarge Scale Visual Recognition Challenge（ＩＬＳＶＲＣ）、物体輪郭検出では、Berkeley Segmentation Dataset（ＢＳＤ５００）である。また、Microsoft（登録商標）

COCOでは、物体ラベルとその物体領域との両方のアノテーションが付与されているため、物体検出タスクと物体輪郭検出タスクとの両方を同時に学習させるためにこのデータセットを用いても良い。また、エッジ検出では、正解ラベルを人間が与えて学習用データを作成しても良いが、Ｓｏｂｅｌ法やＣａｎｎｙ法のような既存アルゴリズムの出力を正解ラベルとして利用し、それらを模倣するようにネットワークを学習させても良い。

【００７４】

以上のように本実施形態によれば、エッジ検出、物体輪郭検出、物体検出という空間的スケールの異なるマルチタスクを行うＣＮＮのニューラルネットワークを実現できる。これにより、画像に対しエッジ検出、物体輪郭検出、イベントセグメント検出の各タスクが一度に処理できるようになる。

【００７５】

（第３の実施形態）

本実施形態では、問題の複雑さの尺度が異なるマルチタスクによりカメラの撮像制御用および現像用の画像認識を実現するＣＮＮを例に説明する。本実施形態で扱う識別タスクは、具体的には合焦判定、笑顔検出、色温度推定、シーン分類とする。

【００７６】

ここで、合焦判定は、カメラのコントラスト方式のオートフォーカス制御で行う合焦判定であり、低次の特徴で実現できる。笑顔検出は、一般に口角の形に基づいて行われるため、中間程度のスケールの特徴が有効であると考えられる。色温度推定は、画像撮影環境における光源の色味による被写体の色変化を軽減させるホワイトバランス補正に利用される。色温度は無彩色領域における色の偏りを元に推定されるが、プリミティブな特徴である色と、画像中の無彩色領域を見つけるより高次な特徴とが複合的に有効なことが考えられるので、複数の中間層の出力を元に識別する識別ユニットを考える。シーン分類では、「ポートレイト」、「記念撮影」、「風景」、「夜景」、「市街地」、「マクロ」等のセンサ感度や絞り・シャッター速度等の制御パラメータを決める際の参考になる、撮影シーンの分類を行う。これは画像全体の情報が必要になるため最終層で実行する。以上から各タスクの空間的スケールは、合焦判定、笑顔判定、シーン分類の順に大きくなり、色温度推定は複数のスケールの特徴量を用いる。

【００７７】

図４（ｃ）は、本実施形態の画像認識装置４５０の機能構成例を示すブロック図である。なお、本実施形態に係る画像認識装置４５０のハードウェア構成については図３と同様である。

本実施形態の画像認識装置４５０は、画像取得部４５１、前処理部４５２、第１中間特徴量取得部４５３、第２中間特徴量取得部４５４、第３中間特徴量取得部４５５、合焦判定部４５６、笑顔検出部４５７、色温度推定部４５８、シーン分類部４５９を有する。画像認識装置４５０が有するこれらの機能の詳細については、図５（ｃ）等を用いて後述する。

【００７８】

図５（ｃ）は、本実施形態における画像認識装置４５０による処理手順の一例を示すフローチャートである。本実施形態では、第２の実施形態と同様に図１４に示したような特徴抽出ユニットおよび識別ユニットを採用し、基本的な処理の流れは第２の実施形態と同様であることから、図１６を参照しながら第２の実施形態との差異を中心に説明する。

【００７９】

まず、Ｓ５５１において、画像取得部４５１は、撮像デバイスから画像を取得する。次に、Ｓ５５２においては、前処理部４５２は、Ｓ５５１で取得した画像に対する前処理を実行する。この処理は第２の実施形態で説明したＳ５３２と同様である。

【００８０】

次に、Ｓ５５３においては、第１中間特徴量取得部４５３は、前処理がなされた入力画像から第１中間特徴量を取得する。ここで、第１中間特徴量取得部４５３は第２の実施形態と同様に図１６に示すようなＣｏｎｖ×２＋Ｐｏｏｌの特徴抽出ユニット１６０２によ

り実現され、合焦判定に係る特徴量を入力画像１６０１から取得する。そして、Ｓ５５４においては、合焦判定部４５６は、第１中間特徴量に基づいて画像が合焦か否かを判定する。この判定は２値分類であるが、第２の実施形態でのエッジ検出と異なり、特徴抽出ユニット１６０２のレスポンスマップ全体と識別ユニットのＦＣ層とが全結合され、識別ユニット１６０５で画像全体に対し２値の合焦か非合焦かが判定される。

【００８１】

次に、Ｓ５５５においては、第２中間特徴量取得部４５４は、第１中間特徴量に基づき、第２中間特徴量を取得する。ここで、第２中間特徴量取得部４５４は第２の実施形態と同様にＣｏｎｖ×３＋Ｐｏｏｌの特徴抽出ユニット１６０３により実現され、笑顔検出に係る特徴量を取得する。そして、Ｓ５５６においては、笑顔検出部４５７は、第２実施形態での物体輪郭検出と同様に、２つめの特徴抽出ユニット１６０３のレスポンスマップの各点で、識別ユニット１６０６により笑顔の有無を２値判定する。

【００８２】

次に、Ｓ５５７においては、第３中間特徴量取得部４５５は、第２中間特徴量に基づき、第３中間特徴量を取得する。ここで、第３中間特徴量取得部４５５は第２の実施形態と同様にＣｏｎｖ×４＋Ｐｏｏｌの特徴抽出ユニット１６０４により実現され、撮影シーンに係る特徴量を取得する。

【００８３】

Ｓ５５８においては、色温度推定部４５８は、第１中間特徴量および第３中間特徴量に基づき色温度の推定を行う。本工程に係る識別ユニット１６０８は複数の層から入力を受け取るため、ＦＣ層が１つめの特徴抽出層と３つめの特徴抽出層と全結合している。また色温度は連続値であるため、分類ではなく回帰を行い、最終層は線形写像とする。

【００８４】

次にＳ５５９においては、シーン分類部４５９は、第３中間特徴量から画像の撮影シーンを分類する。この処理では、第２の実施形態における物体検出と同様に、画像全体に対してマルチクラスの分類を行う。そのため最終層はＳｏｆｔｍａｘの識別ユニット１６０７を用いる。

【００８５】

以上の認識時の処理により、図１６に示すようなネットワーク構造によって、入力画像に対し、合焦判定、笑顔検出、色温度推定、シーン分類の４つのタスクを、１つのネットワークの処理で同時に実現できる。次に、本実施形態におけるＳ５５３～Ｓ５５９で利用する特徴抽出ユニットおよび識別ユニットの学習方法と、図１６に示したネットワーク構造の探索方法に関し、第２の実施形態と差異の部分について説明する。

【００８６】

ネットワーク構造の探索では、まず複数の中間層の出力を用いる色温度推定は除き、合焦判定、笑顔検出、シーン分類の３つのタスクについて、第２の実施形態と同様の方法を用いる。その後、色温度推定の識別層のパラメータのみを学習パラメータとして、色温度推定がより高精度になる中間層を探索する。その際は、ネットワーク構造の探索の結果、全体でｎ個の中間ユニットからなる構造になった場合、ｎ個中の２つの組み合わせを探索すれば２つの中間層の入力を用いるパターンの全ての組み合わせを網羅できる。また、色温度推定は回帰の問題となるため、識別ユニットの最終層は線形写像になる。そして２乗誤差のロス関数を用いてパラメータの最適化を行う。その他の学習方法に関しては第２の実施形態と同様の処理で実現できるため説明を省略する。

【００８７】

以上のように本実施形態によれば、合焦判定、笑顔検出、色温度推定、シーン分類というカメラ制御時に有用なマルチタスクを行うＣＮＮのニューラルネットワークを実現できる。これにより、認識時の処理によって各タスクが一度に処理できるようになる。

【００８８】

（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

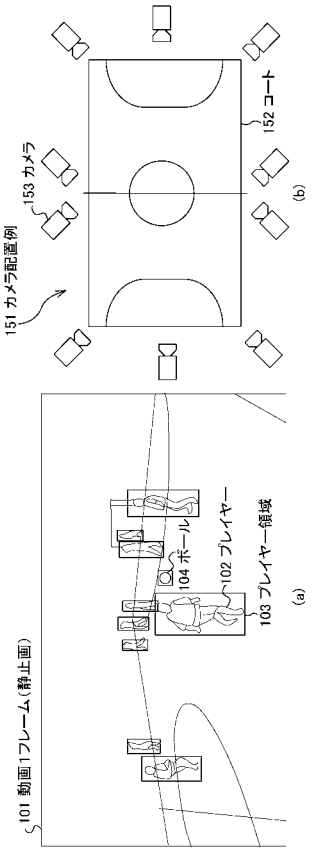
記憶媒体を介してシステム又は装置に供給し、そのシステム又は装置のコンピュータにおける１つ以上のプロセッサーがプログラムを読出し実行する処理でも実現可能である。また、１以上の機能を実現する回路（例えば、ＡＳＩＣ）によっても実現可能である。

【符号の説明】

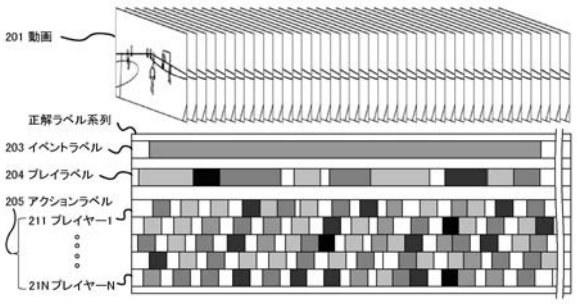
【００８９】

４０５　第１中間特徴量取得部
４０６　第２中間特徴量取得部
４０７　第３中間特徴量取得部
４０８　アクション識別部
４０９　プレイ識別部
４１０　イベントセグメント識別部

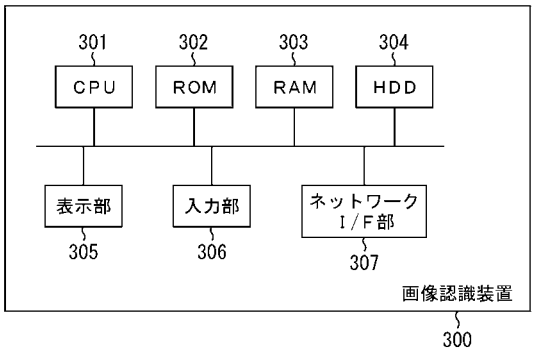
【図１】



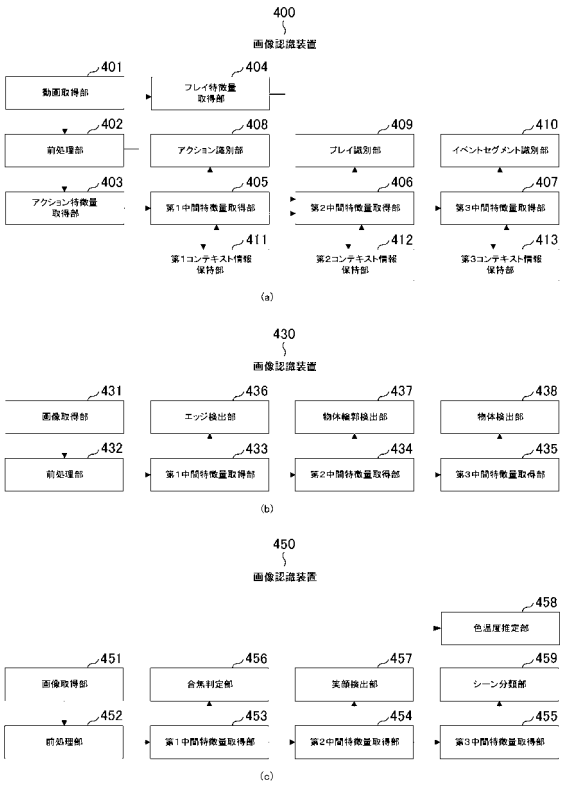
【図２】



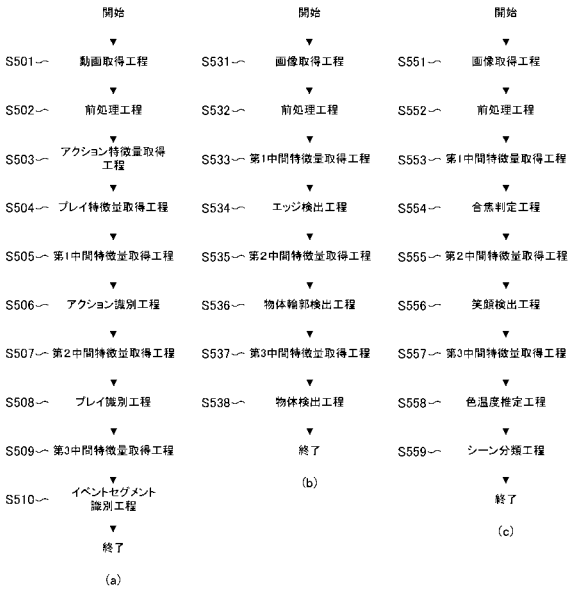
【図３】


301
302
303
304
CPU
ROM
RAM
HDD
表示部
入力部
ネットワーク
I/F部
305
306
307
画像認識装置
300

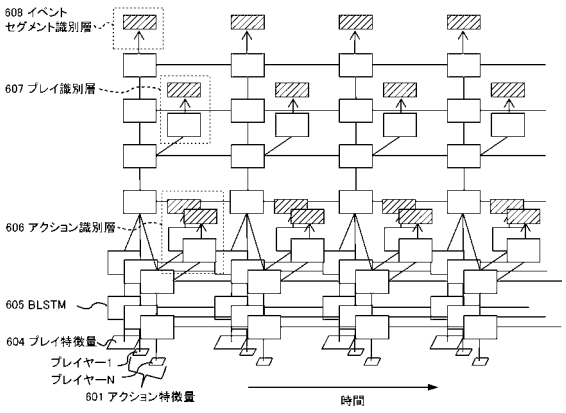
【図４】


400
画像認識装置
401
動画取得部
404
プレイ特徴量取得部
402
前処理部
408
アクション識別部
409
プレイ識別部
410
イベントセグメント識別部
403
アクション特徴量取得部
405
第1中間特徴量取得部
406
第2中間特徴量取得部
407
第3中間特徴量取得部
411
第1コンテキスト情報保持部
412
第2コンテキスト情報保持部
413
第3コンテキスト情報保持部
(a)
430
画像認識装置
431
画像取得部
436
エッジ検出部
437
物体輪郭検出部
438
物体検出部
432
前処理部
433
第1中間特徴量取得部
434
第2中間特徴量取得部
435
第3中間特徴量取得部
(b)
450
画像認識装置
458
色温度推定部
451
画像取得部
456
合焦判定部
457
笑顔検出部
459
シーン分類部
452
前処理部
453
第1中間特徴量取得部
454
第2中間特徴量取得部
455
第3中間特徴量取得部
(c)

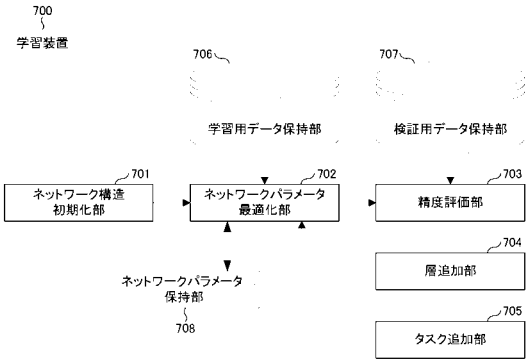
【図５】


開始
S501 動画取得工程
S502 前処理工程
S503 アクション特徴量取得工程
S504 プレイ特徴量取得工程
S505 第1中間特徴量取得工程
S506 アクション識別工程
S507 第2中間特徴量取得工程
S508 プレイ識別工程
S509 第3中間特徴量取得工程
S510 イベントセグメント識別工程
終了
(a)
開始
S531 画像取得工程
S532 前処理工程
S533 第1中間特徴量取得工程
S534 エッジ検出工程
S535 第2中間特徴量取得工程
S536 物体輪郭検出工程
S537 第3中間特徴量取得工程
S538 物体検出工程
終了
(b)
開始
S551 画像取得工程
S552 前処理工程
S553 第1中間特徴量取得工程
S554 合焦判定工程
S555 第2中間特徴量取得工程
S556 笑顔検出工程
S557 第3中間特徴量取得工程
S558 色温度推定工程
S559 シーン分類工程
終了
(c)

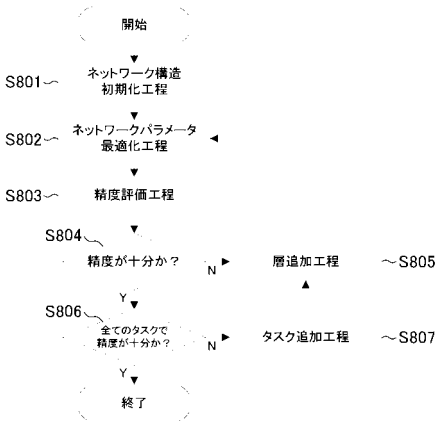
【図６】


609 イベントセグメント識別層
607 プレイ識別層
606 アクション識別層
605 BLSTM
604 プレイ特徴量
プレイヤー1
プレイヤーN
601 アクション特徴量
時間

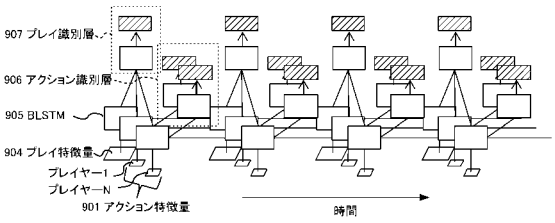
【図７】


700
学習装置
706
707
学習用データ保持部
検証用データ保持部
701
ネットワーク構造
初期化部
702
ネットワークパラメータ
最適化部
703
精度評価部
704
層追加部
705
タスク追加部
ネットワークパラメータ
保持部
708

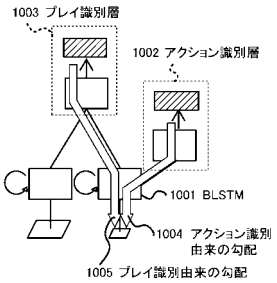
【図８】


開始
S801
ネットワーク構造
初期化工程
S802
ネットワークパラメータ
最適化工程
S803
精度評価工程
S804
精度が十分か？
N
層追加工程
～S805
Y
S806
全てのタスクで
精度が十分か？
N
タスク追加工程
～S807
Y
終了

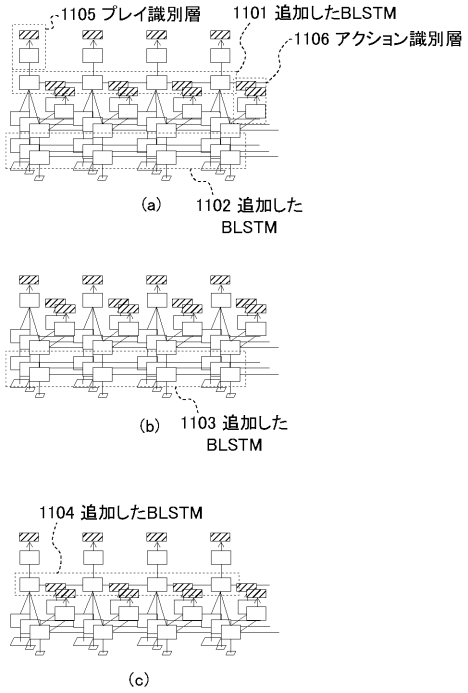
【図９】


907 プレイ識別層
906 アクション識別層
905 BLSTM
904 プレイ特徴量
プレイヤー1
プレイヤーN
901 アクション特徴量
時間

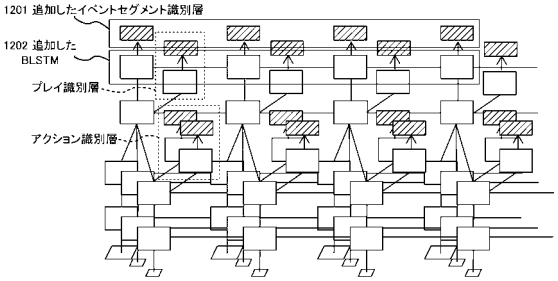
【図１０】


1003 プレイ識別層
1002 アクション識別層
1001 BLSTM
1004 アクション識別
由来の勾配
1005 プレイ識別由来の勾配

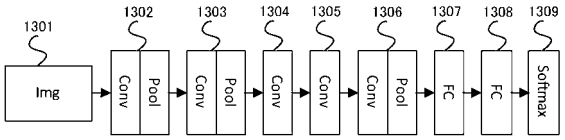
【図１１】



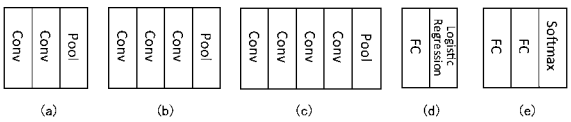
【図１２】



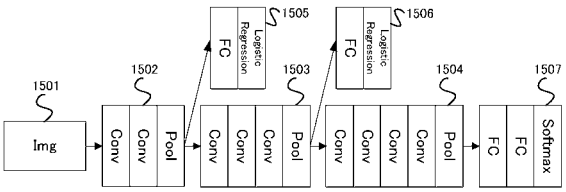
【図１３】



【図１４】



【図１５】



【図１６】