(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第6922005号
(P6922005)

(45)発行日　令和3年8月18日(2021.8.18)　(24)登録日　令和3年7月30日(2021.7.30)

(51)Int.Cl.			F I		
G06T	7/00	(2017.01)	G06T	7/00	350C
G06T	7/11	(2017.01)	G06T	7/11	
G06N	3/04	(2006.01)	G06N	3/04	

請求項の数 19　外国語出願　(全 27 頁)

(21)出願番号　特願2020-7450(P2020-7450)
(22)出願日　令和2年1月21日(2020.1.21)
(65)公開番号　特開2020-119568(P2020-119568A)
(43)公開日　令和2年8月6日(2020.8.6)
審査請求日　令和2年3月2日(2020.3.2)
(31)優先権主張番号　1900883.8
(32)優先日　平成31年1月22日(2019.1.22)
(33)優先権主張国・地域又は機関
英国(GB)

(73)特許権者　000003078
株式会社東芝
東京都港区芝浦一丁目１番１号
(74)代理人　100108855
弁理士　蔵田　昌俊
(74)代理人　100103034
弁理士　野河　信久
(74)代理人　100179062
弁理士　井上　正
(74)代理人　100075672
弁理士　峰　隆司
(74)代理人　100153051
弁理士　河野　直樹
(74)代理人　100162570
弁理士　金子　早苗

最終頁に続く

(54)【発明の名称】コンピュータビジョンのシステムおよび方法

(57)【特許請求の範囲】
【請求項１】
画像を受け取ることと、
第１の特徴マップを作るために共通の処理ステージを用いて前記画像を処理することと、
前記第１の特徴マップを受け取る第１および第２の並列分岐を備える並列処理ステージに前記第１の特徴マップを入力することと、
セマンティックセグメント化された画像を作るために前記第１および第２の分岐の出力を融合ステージで結合することと、
を備え、
前記融合ステージは、前記第１の分岐の前記出力をアップサンプリングすることと、前記第１の分岐の前記アップサンプリングされた出力を前記第２の分岐の前記出力に加算することと、を備え、前記第１の分岐の前記アップサンプリングされた出力は、前記第２の分岐の前記出力への加算の前に深さ方向畳み込みを受け、
前記共通の処理ステージは、ニューラルネットワークを備え、前記ニューラルネットワークは、第１の特徴マップを作るために分離可能な畳み込みを実行し、前記画像をダウンサンプリングするように構成された少なくとも１つの分離可能な畳み込みモジュールを有し、前記第１の分岐は、分離可能な畳み込みを実行するように構成された少なくとも１つの分離可能な畳み込みモジュールを備えるニューラルネットワークを備える、
画像をセグメント化する画像処理方法。

【請求項２】
第２の分岐は、前記第１の特徴マップが前記第１の分岐の前記出力と結合されることを可能にするためにスキップ接続として機能する、請求項１に記載の画像処理方法。
【請求項３】
前記第１および第２の分岐からの前記出力は、１つのステージのみで結合される、請求項１または２に記載の画像処理方法。
【請求項４】
前記第１の分岐内の層の数は、前記共通の処理ステージ内の層の数より多い、請求項１から３のいずれかに記載の画像処理方法。
【請求項５】
前記加算することは、アップサンプリングされ、深さ方向畳み込みされた前記第１の分岐の出力と前記第２の分岐との対応する値を加算することを備える、請求項１から４のいずれかに記載の画像処理方法。
【請求項６】
前記融合ステージにおける前記深さ方向畳み込みは、１より大きい膨張係数を用いて実行される、請求項１に記載の画像処理方法。
【請求項７】
前記第１の分岐のアップサンプリングされた出力は、１より大きい膨張係数を有する深さ方向畳み込みを受け、前記第２の分岐の前記出力は、加算の前に２次元畳み込みを受ける、請求項６に記載の画像処理方法。
【請求項８】
前記第１および第２の分岐の前記結合された出力は、分類器によって処理される、請求項１から７のいずれかに記載の画像処理方法。
【請求項９】
前記第１の分岐のステージ内の前記分離可能な畳み込みモジュールは、深さ方向畳み込みモジュールである、請求項１から８のいずれかに記載の画像処理方法。
【請求項１０】
前記第１の分岐内の前記分離可能な畳み込みモジュールは、深さ方向の分離可能な畳み込みモジュールである、請求項１から９のいずれかに記載の画像処理方法。
【請求項１１】
前記第１の分岐内の前記分離可能な畳み込みモジュールは、ボトルネックアーキテクチャモジュールである、請求項１から１０のいずれかに記載の画像処理方法。
【請求項１２】
複数のボトルネック残差アーキテクチャモジュールが存在する、請求項１１に記載の画像処理方法。
【請求項１３】
前記複数のボトルネック残差アーキテクチャモジュールの後にピラミッドプーリングモジュールが設けられる、請求項１２に記載の画像処理方法。
【請求項１４】
モデルをトレーニングする方法であって、前記モデルは、画像をセマンティックセグメント化するモデルであり、前記モデルは、
第１の特徴マップを作るための共通の処理ステージと、
前記第１の特徴マップを受け取る第１および第２の並列分岐を備える並列処理ステージと、
セマンティックセグメント化された画像を作るために前記第１および第２の分岐の出力を結合する融合ステージと、
を備え、
前記融合ステージは、前記第１の分岐の前記出力をアップサンプリングすることと、前記第１の分岐の前記アップサンプリングされた出力を前記第２の分岐の前記出力に加算することと、を備え、前記第１の分岐の前記アップサンプリングされた出力は、前記第２の

分岐の前記出力への加算の前に深さ方向畳み込みを受け、
前記共通の処理ステージはニューラルネットワークを備え、前記ニューラルネットワークは、第１の特徴マップを作るために、分離可能な畳み込みを実行し、前記画像をダウンサンプリングするように構成された少なくとも１つの分離可能な畳み込みモジュールを有し、前記第１の分岐は、分離可能な畳み込みを実行するように構成された少なくとも１つの分離可能な畳み込みモジュールを備えるニューラルネットワークを備え、
前記トレーニングする方法は、
トレーニングデータを提供することと、ここで前記トレーニングデータは、画像および前記画像に関するセマンティックセグメント化された情報を備え、
入力として前記画像を使用し、出力として前記セマンティックセグメント化された情報を使用して前記モデルをトレーニングすることと、ここで前記共通の処理ステージおよび前記並列処理ステージは一緒にトレーニングされる、
を含む、方法。
【請求項１５】
トレーニング中にフィルタの個数を適応させることと、より重要性の低いフィルタを破棄するためにその個数を減らすこととをさらに備える、請求項１４に記載の方法。
【請求項１６】
前記共通の処理ステージおよび／または第１の処理分岐ステージに対して少なくとも１つの追加出力を加えるためにトレーニング中に前記モデルを適応させることをさらに備え、前記方法は、入力として前記画像を使用してトレーニングすることと、前記出力と前記少なくとも１つの追加出力との両方での前記セマンティックセグメント化された情報の両方との比較によって損失を決定することと、両方の出力からの前記決定された損失を使用することによってトレーニング中に重みを更新することとをさらに備える、請求項１４に記載の方法。
【請求項１７】
インターフェースとプロセッサと
を備え、
前記インターフェースは、画像入力を有し、画像を受け取るように適応され、
前記プロセッサは、
第１の特徴マップを作るために共通の処理ステージを用いて前記画像を処理し、
前記第１の特徴マップを受け取る第１および第２の並列分岐を備える並列処理ステージに前記第１の特徴マップを入力し、
セマンティックセグメント化された画像を作るために前記第１および第２の分岐の出力を融合ステージで結合するように適応され、
前記融合ステージは、前記第１の分岐の前記出力をアップサンプリングすることと、前記第１の分岐の前記アップサンプリングされた出力を前記第２の分岐の前記出力に加算することと、を備え、前記第１の分岐の前記アップサンプリングされた出力は、前記第２の分岐の前記出力への加算の前に深さ方向畳み込みを受け、
前記共通の処理ステージは、ニューラルネットワークを備え、前記ニューラルネットワークは、第１の特徴マップを作るために分離可能な畳み込みを実行し、前記画像をダウンサンプリングするように構成された少なくとも１つの分離可能な畳み込みモジュールを有し、前記第１の分岐は、分離可能な畳み込みを実行するように構成された少なくとも１つの分離可能な畳み込みモジュールを備えるニューラルネットワークを備える、
画像をセグメント化する画像処理システム。
【請求項１８】
車両の検出システムであって、前記検出システムは、画像を受け取り、前記画像をセグメント化することによって前記画像から物体を決定するように適応された、請求項１７の画像処理システムを含む、検出システム。
【請求項１９】
請求項１から１６のいずれかの方法をコンピュータに実行させるように適応されたコン

ピュータ可読命令を担持する非一時的キャリアメディア。
【発明の詳細な説明】
【技術分野】
【０００１】
諸実施形態はコンピュータビジョンのシステムおよび方法に関する。
【背景技術】
【０００２】
多数のコンピュータビジョンタスク、たとえば物体の認識およびレジストレーションは、画像の領域にラベルが与えられる、画像のセグメンテーションを必要とする。セマンティックセグメンテーションは、詳細な画素レベルの仕様を提供し、これは、障害物検出および正確な境界検出をしばしば必要とするアプリケーションに特に適する。そのようなアプリケーションは、自律車両および運転者支援、組み込みデバイス、ならびにウェアラブルデバイスを含むが、これに限定はされない。
【０００３】
近代のセマンティックセグメンテーション方法は、非常に正確な結果を達成するが、しばしば、効率の低下という犠牲を払う。畳み込みニューラルネットワーク（ＣＮＮ）の最近の開発は、これらのネットワークによって達成される結果の顕著な改善を示す。しかし、これらの有効性は、モデルに含まれるオペレーションおよびパラメータの数に大きく依存する。最近のセマンティックセグメンテーション方法は、処理がハイエンドグラフィックス処理ユニット（ＧＰＵ）上で実行される場合であっても、単一の画像の物体分類を実行するのに１秒超を要する。これらの方法の複雑さは、リアルタイムアプリケーションでのそれらの展開を妨げる。
【０００４】
自律運転は、複雑なタスクであり、物体の検出および分類は、他のタイムクリティカルなタスクの前処理ステップにすぎない。そのようなシステムは、しばしば、リアルタイムよりも高速の能力を有するシステムから利益を得る。したがって、物体の検出および分類システムが、物体分類の精度を損なうことなく、リアルタイムよりも高速の性能が可能である、セマンティックセグメンテーションの新しい手法が必要である。
【図面の簡単な説明】
【０００５】
【図１】車両で実施される実施形態に係るシステムの図。
【図２】図１のシステム内で使用され得るニューラルネットワークの簡潔なフロー図。
【図３】物体分類用の高速ニューラルネットワークの詳細なフロー図。
【図４】一実施形態によるネットワークアーキテクチャの概略図。
【図５】一実施形態によるピラミッドプーリングモジュールアーキテクチャの概略図。
【図６】物体分類用のニューラルネットワークの例のトレーニングプロセスのフロー図。
【図７】パラメータ枝刈りプロセスを示すフロー図。
【図８】一実施形態による都市景観に対するトレーニング曲線の図。
【図９】都市景観妥当性検査セットに関する視覚的比較の図。第１列は入力ＲＧＢ画像、第２の列はグラウンドトゥルースラベル、最後の列はＦａｓｔ－ＳＣＮＮの出力である。
【図１０】Ｆａｓｔ－ＳＣＮＮのセグメンテーション結果の視覚的比較の図。第１列は入力ＲＧＢ画像、第２の列はＦａｓｔ－ＳＣＮＮの出力、最後の列はスキップ接続の寄与を０にした後のＦａｓｔ－ＳＣＮＮの出力である。
【図１１】一実施形態に従って使用され得るデバイスの概略図。
【発明を実施するための形態】
【０００６】
一実施形態では、
画像を受け取ることと、
第１の特徴マップを作るために共通の処理ステージを用いて前記画像を処理することと、

前記第１の特徴マップを並列処理ステージに入力することと、ここで前記第２の処理ステージは、第１の特徴マップを受け取る第１および第２の並列分岐を備え、
セマンティックセグメント化された画像を作るために第１および第２の分岐の出力を結合することとを備え、
共通の処理ステージは、ニューラルネットワークを備え、ニューラルネットワークは、第１の特徴マップを作るために分離可能（separable）な畳み込みを実行し、画像をダウンサンプリングするように構成された少なくとも１つの分離可能な畳み込みモジュールを有し、前記第１の分岐は、分離可能な畳み込みを実行するように構成された少なくとも１つの分離可能な畳み込みモジュールを備えるニューラルネットワークを備える、
画像をセグメント化する画像処理方法が提供される。

【０００７】
開示されるシステムは、コンピュータ技術に結び付けられ、コンピューティングの領域から生じる技術的問題、すなわち、リアルタイムセマンティックセグメンテーションの提供という技術的問題に対処する。開示されるシステムは、２つの並列分岐を有する並列処理ステージに先行する共通の処理ステージによって、この技術的問題を解決する。共通の処理ステージは、画像をダウンサンプリングし、低レベルの特徴を計算する。したがって、共通の処理ステージは２つの特徴をサービスし、それにより低レベルの特徴が抽出されることを可能にし、また、画像からコンテキスト情報を抽出するように構成された分岐に入力することを可能にするために、画像をダウンサンプリングする。

【０００８】
上記方法は、リアルタイムセマンティックセグメンテーションを可能にし、実際に、高解像度画像データ（１０２４×２０４８画素）に対してリアルタイムセマンティックセグメンテーションを超える結果を生み出すことができる。したがって、上記方法は、少メモリ組み込みデバイス上での効率的な計算に適する。上の実施形態は、高速セグメンテーション用の２分岐方法を使用する。第１の「共通の」処理ステージは、多重解像度分岐の浅い低レベル特徴（shallow low level feature）を同時に計算する「ダウンサンプリングを学習する」モジュールと考えられ得る。

【０００９】
このアーキテクチャは、高解像度での空間詳細とより低い解像度で抽出される深い特徴（deep feature）との組合せを可能にする。

【００１０】
上の実施形態では、第１の分岐は、画像またはシーンを構成する物体を分類するか他の形で識別することを目指し、シーン内に存在する物体にラベルを割り当てる。対して、第２の分岐は、主に、高解像度セグメンテーション結果の物体境界に関する情報を維持する処理を担う。

【００１１】
上の実施形態では、第１の分岐は、深層ネットワークを備える。

【００１２】
上記の実施形態のモデルは、以下の理由から計算コストを削減する。
ｉ）コンテキスト処理では、より複雑で抽象的な特徴を学習する必要があり、したがってより深いネットワークが必要である。しかし、より高解像度の入力は不要であり、ゆえにコンテキスト分岐上でより低解像度の入力を使用することによってモデルコストが低減される。
ｉｉ）境界処理では、高解像度入力の必要がある。しかし、大きい視野を見る必要はなく、ゆえに詳細分岐内で少数の層を使用することによってモデルコストが低減される。
ｉｉｉ）したがって、上で述べたように２つの異なる目的のために２つの分岐を動作させることは、全体的なモデルコストを下げる。
ｉｖ）第２の分岐は、スキップ接続と考えられることができ、発明者らは、第１および第２の分岐が、初期処理ステージを共有できることを理解した。
ｖ）画像処理ネットワークが、少容量になるように設計された。

【００１３】
一実施形態では、第２の分岐は、第１の特徴マップが第１の分岐の出力と結合されることを可能にするためにスキップ接続として機能する。いくつかの実施形態では、特徴マップは、未変更で第２の分岐を通過される。他の実施形態では、第１の特徴マップは、たとえば、第１の分岐の出力と結合される前に第１の特徴マップの解像度を変更するために、１つまたは複数のニューラルネットワーク層を通過される。一実施形態では、第２の分岐は、１つまたは複数の２Ｄ畳み込み層を備える。
【００１４】
一実施形態では、第１および第２の分岐からの出力は、１つのステージのみで結合される。一実施形態では、２つの分岐だけがある。
【００１５】
第１の分岐は、コンテキスト情報が抽出されることを可能にする複数のモジュールを備え、第１の分岐は、深層ネットワークを備え、第１の分岐内の層の個数は、共通処理ステージ内の層の個数より多い。
【００１６】
一実施形態では、上記システムは、深さ方向（depth-wise）の分離可能な畳み込みを使用して、標準的な畳み込み（Ｃｏｎｖ２ｄ）を、空間畳み込みまたはチャネル方向（channel-wise）畳み込みとしても知られる深さ方向の畳み込み（ＤＷＣｏｎｖ）とそれに続く１×１点方向（1 x 1 point-wise）畳み込み層とに因数分解する。したがって、チャネル間（cross-channel）相関および空間相関は、独立に計算され、これは、パラメータの数を劇的に低減し、より少数の浮動小数点演算と高速の実行時間とをもたらす。したがって、計算コストおよびメモリ要件が削減される。
【００１７】
一実施形態では、第１および第２の分岐の出力は、１つのステージのみで結合される。第１および第２の分岐の出力は、１回だけ結合される。
【００１８】
一実施形態では、第１および第２の分岐の出力は、融合ステージで結合され、前記融合ステージは、第１の分岐の出力をアップサンプリングすることと、第１の分岐のアップサンプリングされた出力を第２の分岐の出力に加算することとを備え、加算することは、第１の分岐のアップサンプリングされた出力と第２の分岐との対応する値を加算することを備える。
【００１９】
さらなる実施形態では、第１および第２の分岐の出力は、加算によって結合される。一実施形態では、サイズａ＊ｂ＊ｃの第２の特徴マップ（アップサンプリングの後の低解像度から）は、サイズａ＊ｂ＊ｃの結合された特徴マップＣを作成するために、両方の数を加算することによってサイズａ＊ｂ＊ｃの第２の特徴マップと結合される。連結（concatenation）は不要であり、これは、メモリフットプリントを小さく保つことを可能にする。
【００２０】
第１の分岐からのアップサンプリングされた出力は、第２の分岐からの出力への加算の前に深さ方向畳み込みを受け、ここで深さ方向畳み込みは、１より大きい膨張係数（dilation factor）を用いて実行される。
【００２１】
さらなる実施形態では、１より大きい膨張係数を有する深さ方向畳み込みを受けた第１の分岐からのアップサンプリングされた出力、および第２の分岐からの出力は、加算の前に２次元畳み込みを受ける。
【００２２】
一実施形態では、第１の分岐内の深さ方向畳み込みモジュールは、ボトルネックアーキテクチャモジュールである。ボトルネックアーキテクチャモジュールは、特徴拡張（feature expansion）を実施し、その後に深さ方向畳み込みが行われ、その後に点畳み込みが

行われる。したがって、これらはまた、深さ方向の分離可能な畳み込みを実施しているとも考えられ得る。

【００２３】

一実施形態では、第１の分岐の深さ方向畳み込みモジュールは、ボトルネック残差アーキテクチャモジュールである。複数の残差ボトルネックアーキテクチャモジュールが存在し得る。残差ボトルネックアーキテクチャモジュールからの出力チャネルの個数は、複数の残差ボトルネックアーキテクチャモジュール内の後続モジュールでは増加し得る。

【００２４】

第１の特徴マップは、前記深さ方向畳み込みモジュールによって処理される前に、標準的な畳み込みを受け得る。最終的な標準畳み込みモジュールは、残差ボトルネックアーキテクチャモジュールの後に設けられ得る。

【００２５】

一実施形態では、ピラミッドプーリングモジュールが、最終的な標準畳み込みモジュールと残差ボトルネックアーキテクチャモジュールとの間に設けられる。

【００２６】

共通の処理ステージは、複数の深さ方向の分離可能畳み込みモジュールを備えることができる。画像は、共通の処理ステージ内の前記深さ方向の分離可能畳み込みモジュールによって処理される前に、標準的な畳み込みを受け得る。

【００２７】

一実施形態では、共通の処理ステージは、低レベル特徴の共有が有効であり、効率的に実施されることを保証するために、３つの層を備える。一実施形態では、第１の層は、標準的な畳み込み層（Ｃｏｎｖ２Ｄ）であり、残りの２つの層は、深さ方向の分離可能な畳み込み層（ＤＳＣｏｎｖ）である。ＤＳＣｏｎｖは、計算的により効率的であるが、入力画像だけが３つのチャネルを有し、ＤＳＣｏｎｖの計算的な影響をこのステージで重大なものにするので、Ｃｏｎｖ２Ｄがここで使用される。

【００２８】

一実施形態では、共通の処理ステージの層は、ストライド２を使用し、これに、バッチ正規化とＲｅＬＵとが続く。畳み込み層および深さ方向層の空間カーネルサイズは、３×３である。一実施形態では、深さ方向畳み込みと点方向畳み込みとの間の非線形性が、省略される。

【００２９】

一実施形態では、第１および第２の分岐の結合された出力は、分類器によって処理される。一実施形態では、分類器は、少なくとも１つの分離可能な畳み込みモジュールを備える。分類器は、ｓｏｆｔｍａｘ層を備えることができる。少なくとも１つの分離可能な畳み込みモジュールは、深さ方向の分離可能な畳み込みモジュールである。

【００３０】

一実施形態では、浮動小数点乗算が、整数演算または２進演算と比較して高コストなので、ランタイムは、ＤＣＮＮフィルタおよびアクティブ化値に関する量子化技法を使用して、さらに削減され得る。

【００３１】

一実施形態では、２進量子化技法が使用され得る。

【００３２】

さらに、一実施形態では、事前にトレーニングされたネットワークのサイズを低減するために枝刈り（pruning）が適用され、その結果、より高速のランタイムと、より小さいパラメータセットと、より小さいメモリフットプリントとをもたらす。

【００３３】

一実施形態では、モデルをトレーニングする方法が提供され、前記モデルは、画像をセマンティックセグメント化するモデルであり、モデルは、

第１の特徴マップを作るための共通の処理ステージと、

並列処理ステージと、ここで前記第２の処理ステージは、第１の特徴マップを受け取る

第１および第２の並列分岐を備え、
セマンティックセグメント化された画像を作るために第１および第２の分岐の出力を結合することと
を備え、共通の処理ステージはニューラルネットワークを備え、ニューラルネットワークは、第１の特徴マップを作るために、分離可能な畳み込みを実行し、画像をダウンサンプリングするように構成された少なくとも１つの分離可能な畳み込みモジュールを有し、前記第１の分岐は、分離可能な畳み込みを実行するように構成された少なくとも１つの分離可能な畳み込みモジュールを備えるニューラルネットワークを備え、
トレーニング方法は、
トレーニングデータを提供することと、ここでトレーニングデータは、画像および前記画像に関するセマンティックセグメント化された情報を備え、
入力として前記画像を使用し、出力としてセマンティックセグメント化された情報を使用して前記モデルをトレーニングすることとを備え、ここで２つのステージは一緒にトレーニングされる。

【００３４】

さらなる実施形態では、パラメータ枝刈りが、トレーニング中に実施され得る。

【００３５】

さらなる実施形態では、トレーニング方法は、前記第１の分岐に対して第２の出力を加えるためにトレーニング中にモデルを適応させることをさらに備え、方法は、入力として画像を使用してトレーニングすることと、出力と第２の出力との両方でのセマンティックセグメント化された情報の両方との比較によって損失を決定することと、両方の出力からの決定された損失を使用することによってトレーニング中に重みを更新することとをさらに備える。

【００３６】

一実施形態では、
インターフェースとプロセッサと
を備え、
前記インターフェースは、画像入力を有し、第１の画像を受け取るように適応され、前記プロセッサは、
第１の特徴マップを作るために共通の処理ステージを用いて前記画像を処理し、
並列処理ステージに前記第１の特徴マップを入力し、前記第２の処理ステージは、第１の特徴マップを受け取る第１および第２の並列分岐を備え、
セマンティックセグメント化された画像を作るために第１および第２の分岐の出力を結合する
ように適応され、ここで、共通の処理ステージは、ニューラルネットワークを備え、ニューラルネットワークは、第１の特徴マップを作るために分離可能な畳み込みを実行し、画像をダウンサンプリングするように構成された少なくとも１つの分離可能な畳み込みモジュールを有し、前記第１の分岐は、分離可能な畳み込みを実行するように構成された少なくとも１つの分離可能な畳み込みモジュールを備えるニューラルネットワークを備える
画像をセグメント化する画像処理システムが提供される。

【００３７】

一実施形態では、このシステムは車両上で実施され、このシステムがその上で実施され得る車両は、このシステムに入力高解像度画像を供給する１つまたは複数のスチールデジタルカメラおよびまたはビデオデジタルカメラを備えた自律車両と半自律車両とを含むがこれに限定されない。一実施形態では、このシステムは、車両上に配置されたグラフィックス処理ユニットまたは中央処理装置上で実現される。このシステムの目的は、車両の周囲の物体を分類し、車両の最終目的地に向かう車両の安全なナビゲーションを容易にすることである。

【００３８】

したがって、さらなる実施形態では、車両の検出システムであって、前記検出システム

すなわち、上で説明した画像処理システムを備える検出システムは、前記画像を受信し、前記画像をセグメント化することによって前記画像から物体を決定するように適応される、検出システムが提供される。

【００３９】

図１は、システムが車両とともに移動するように車両に搭載されて提供された物体分類ネットワークの実施形態を示すのに使用される。図１は、自動車１の概略を示し、自動車１は、衝突回避システムを備える。衝突回避システムは、４つのカメラ３、５、７、９を備え、カメラ３、５、７、９は、自動車１の各コーナーに設けられる。カメラのそれぞれは、観察可能な世界のより広い広がりをカメラが見ることを可能にする、広い視野（ＦＯＶ）を有する。一実施形態では、各カメラ３、５、７、９は、幅広いパノラマ画像を作る、非常に広角の魚眼レンズを与えられ得る。各カメラからの画像は、別々に作られるが、前部カメラ３および５のＦＯＶは、視界が遮られる自動車１の前の部分内のエリアを残さないように、オーバーラップしてもよい。後部カメラ７および９のＦＯＶも、視界が遮られる自動車１の背後の部分内のエリアを残さないように、オーバーラップしてもよい。

【００４０】

このシステムの一実施形態では、各カメラ３、５、７、９からの画像は、単一の中央処理装置（ＣＰＵ）またはＧＰＵによって別々に処理される。このシステムのさらなる実施形態では、各カメラ３、５、７、９は、別々のＣＰＵまたはＧＰＵを与えられ、この別々のＣＰＵまたはＧＰＵは、画像を処理し、処理された画像を自動車１の中央ＣＰＵに転送する。

【００４１】

上記実施形態は、運転のための自律システムに関係する。しかし、この画像処理方法は、画像のセマンティックセグメンテーションを必要とするすべてのシステム、たとえばウェアラブル技術などにも適用され得る。

【００４２】

図２は、物体の識別と分類とに関する、一実施形態によるシステムのフロー図を示す。

【００４３】

一実施形態では、物体分類システムは、多重分岐アーキテクチャを有する畳み込みニューラルネットワークとエンコーダ－デコーダフレームワークとの組合せを備え、異なる解像度レベルでの初期畳み込みは、分岐によって共有される。

【００４４】

図２には、ニューラルネットワークが２つの分岐を備える実施形態による、多重分岐畳み込みニューラルネットワークのアーキテクチャが示されている。第１の分岐は、シーンの大域（global）コンテキストを取り込む処理を担い、第２の分岐は、空間詳細を用いて大域コンテキストを洗練させる処理を担う。シーンの局所（local）コンテキストは、フル解像度画像から抽出され、したがって、分岐は、少数の畳み込み層を備える。大域コンテキストは、前記分岐内のより多数の畳み込み層を可能にする、より低い画像解像度で取り込まれる。さらに、システムのこの実施形態では、エンコーダ－デコーダフレームワークも使用される。ニューラルネットワークのアーキテクチャ内にスキップ接続を統合することは、浅い低レベルの特徴が２つの分岐について同時に抽出されることを可能にする、２つの分岐の初期層が共有されることを可能にする。スキップ接続は、ランタイム効率のために１回だけ使用される。さらに、高解像度分岐での（すなわち、共有される初期層での）プーリング畳み込み動作の使用は、ネットワークの低解像度分岐のために画像をより低解像度にダウンサンプリングする必要を回避する。スキップ接続は、ネットワークの初期層内で抽出されるシーンの局所コンテキストが、低解像度サブネットワーク分岐によって抽出される大域特徴とマージされることを可能にする。

【００４５】

ステップＳ２０１では、カメラ３、５、７、９が、シーンの画像を取り込み、ここでシーンは、１つまたは複数の物体を備える場合がある。一実施形態では、シーンは、街路シーンである。取り込まれた画像は、物体分類システムに入力され、ここで、シーン内の物

体は、その結果として、ニューラルネットワークによって識別され、ラベルを付与される。

【００４６】

一実施形態では、ステップ２０１の後に、「ダウンサンプリングを学習する」モジュールが設けられる。入力画像は、フル解像度で「ダウンサンプリングを学習する」モジュールに供給される。ニューラルネットワークは、ステップＳ２０３で、シーンの局所コンテキストを抽出し、入力画像のダウンサンプリングされた表現を生成する。

【００４７】

ステップＳ２０３の後に、一実施形態では、画像の経路は、２つの独立のサブネットワーク分岐に分かれる。低解像度の分岐である第１の分岐では、画像のダウンサンプリングされた表現が供給され、ステップＳ２０５で、シーンの大域コンテキストが、畳み込みニューラルネットワークによって抽出される。第２の分岐は、２つの分岐の特徴が結合されるステップＳ２０７で、空間詳細の回復を可能にするスキップ接続として実施される。

【００４８】

図２では、「ダウンサンプリングを学習する」モジュールは、ニューラルネットワークの多重解像度分岐の浅い低レベルの特徴を計算する。一実施形態に従ってこれがどのように達成されるのかは、図３のステップＳ３０３およびＳ３０５を参照して説明される。

【００４９】

入力画像は、３次元行列フォーマットで表現され得る。各画素は、３つのチャネルを備え、各チャネルは、３つの色すなわち赤、緑、および青（ＲＧＢ）のうちの１つの強度に関連する数値を保持する。標準計算では、２Ｄ畳み込みが、ｈ＊ｗ＊ｃフィルタを使用して使用され得、ｈは高さ、ｗは幅、ｃはチャネルである。ｈ＊ｗ＊３（ＲＧＢを有する画像に対する畳み込み）は、ｈ＊ｗ＊１（ＤＷ畳み込み）に類似するので、標準畳み込みが、ここでは申し分ない。通常は、特徴チャネルの個数がＣＮＮでは素早く増加し、したがって、ｈ＊ｗ＊３２／６４／１２８が、めずらしくはないことに留意されたい。一実施形態では、深さ方向畳み込みが、増加する個数のチャネルと共に使用される。

【００５０】

ステップＳ３０３では、畳み込みニューラルネットワークの第１の層は、前の段落で説明したように、より効率的な物体分類を達成するために、２次元プーリング畳み込み層である。プーリング層は、出力される特徴マップの次元が層のストライドに対する相対的な比率だけ減らされる、ダウンサンプリング動作を実行する。層のストライドは、畳み込みカーネルが画像にまたがってスキャンされる時に畳み込みカーネルがどれほど移動されるのかであると考えられ得る。たとえば、ストライド１（１画素）を有するプーリング畳み込み層は、特徴マップの空間次元に影響しないが、ストライド２を有するプーリング畳み込み層は、特徴マップの次元を２倍だけダウンサンプリングする。より小さい特徴マップは、より高速の推論を生じることができるが、犠牲にされた予測正確さを犠牲にする。

【００５１】

この実施形態では、第１の畳み込み層は、入力ボリュームに複数のフィルタまたはカーネルを適用する。各フィルタはシーンと比較され、２次元特徴マップが生成される。特徴マップは、フィルタとシーン内の物体または特徴との間の一致の結果として引き起こされるアクティブ化の空間的配置を表す。すべての特徴マップが、深さ次元に沿って積み重ねられ、出力ボリュームを作る。たとえば、３２個の出力チャネルからなる畳み込みニューラルネットワークは、３２個のフィルタを入力ボリュームに適用し、３２の深さ次元を有する出力ボリュームをレンダリングするために積み重ねられる３２個の特徴マップを生成する。

【００５２】

一実施形態では、物体分類は、複数の深さ方向の分離可能な畳み込み層を利用するニューラルネットワークを用いて実行される。深さ方向畳み込みでは、２次元畳み込みが、各入力チャネルに対して別々に実行され、入力チャネルごとの特徴マップが生成される。すべての特徴マップは、畳み込み手順の終りに一緒に積み重ねられる。たとえば、カメラ３

、５、７、９は、各画素が、１つは赤、１つは緑、１つは青の３つの値のセットを備えるカラーデジタル画像をニューラルネットワークに供給する。画像が深さ方向の分離可能な畳み込み層を用いて処理される時に、２次元畳み込みが、色ごとに別々に実行される。

【００５３】

深さ方向畳み込みとそれに続く点方向畳み込みは、標準畳み込み層と比較して、物体検出の正確さを大幅には低下させない。情報は、まずチャネル方向レベルで計算され、これに、チャネル方向の情報を結合する、より安価な標準畳み込みが続く。畳み込みは、カーネル１×１のみを使用するので、より高速であり、より少数のパラメータを必要とする。

【００５４】

一実施形態による「ダウンサンプリングを学習する」モジュール内の畳み込み層のパラメータが、表１に提示されている。

【表１】

表１

入力	演算子	出力チャネル	繰返し	ストライド
1048 x 2048 x 3	Conv2d	32	1	2
512 x 1024 x 32	DWConv	48	1	2
256 x 512 x 48	Conv2d (1x1)	48	1	1
256 x 512 x 48	DWConv	64	1	2
128 x 256 x 64	Conv2d (1x1)	64	1	1

【００５５】

一実施形態では、プーリング畳み込み層４０３は、フル解像度画像４０１に３２個のフィルタを適用し、３２個の特徴マップを生成する。

【００５６】

ステップＳ３０５では、３×３のカーネルサイズを有する２つの深さ方向の分離可能な畳み込みが使用される。各深さ方向の分離可能な畳み込みブロックは、深さ方向畳み込み層とそれに続く１×１点方向畳み込み層とを備える。図４では、各深さ方向の分離可能な畳み込みの２つの層が、単一のブロックとして表されている。

【００５７】

第１の深さ方向の分離可能な畳み込みブロック４０５は、３２個の入力チャネルと、４８個の出力チャネルと、ストライド２とを有する。

【００５８】

第２の深さ方向の分離可能な畳み込みブロック４０７は、４８個の入力チャネルと、６４個の出力チャネルと、ストライド２とを有する。

【００５９】

したがって、この実施形態では、「ダウンサンプリングを学習する」モジュールは、５つのニューラルネットワーク層を備える。

【００６０】

図２では、第１のサブネットワーク分岐は、適当なラベルを用いてシーン内の物体にラベルを付けると説明された。これが一実施形態に従ってどのように達成されるのかが、図３のステップＳ３０７およびＳ３０９を参照して説明される。

【００６１】

「ダウンサンプリングを学習する」モジュールの出力では、画像次元すなわち長さｈおよび幅ｗが、ｎ分の１に縮小される。したがって、画像は、n^2分の１でダウンサンプリングされる。このシステムの計算時間および物体分類の達成される正確さは、両方とも、因数ｎの値に反比例する。

【００６２】

ｎという因数は、２と３２との間の範囲にわたる可能性があり、「ダウンサンプリング

を学習する」モジュールに備えられるプーリング層の個数によって決定される。

【００６３】

物体分類のためのネットワークの上で説明された実施形態では、入力画像の空間次元は、ｎ＝８分の１に縮小される。したがって、画像は、「ダウンサンプリングを学習する」モジュールの出力で６４分の１でダウンサンプリングされている。

【００６４】

一実施形態では、深さ方向畳み込み層は、ボトルネックブロック内で使用され得る。ボトルネック残差ブロックは、表２に従って、入力を、ｃ個のチャネルから、高さｈ、幅ｗ、拡張係数ｔ、畳み込みタイプのカーネルサイズ／ストライドｓ、および非線形関数ｆを有するｃ’個のチャネルに転送する。

【００６５】

ボトルネックブロックでは、入力ボリュームは、拡張され、その後、深さ方向畳み込み層とそれに続く点方向畳み込みを用いてフィルタリングされる。

【００６６】

一般に、まず、点方向畳み込みが適用される（行１）。その後、深さ方向畳み込みおよび点方向畳み込み（行２および行３が適用される）。

【００６７】

一実施形態では、以下の処理が続く。

Ｃｏｎｖ２ｄ　１／１
（１×１×ｃ×ｔ＊ｃ個のパラメータ）
（ｈ×ｗ×１×１×ｃ×ｔ＊ｃ個の動作）
Ｃｏｎｖ２ｄ　３／ｓ
（３×３×ｔ＊ｃ×ｃ’個のパラメータ）
（ｈ／ｓ×ｗ／ｓ×３×３×ｔ＊ｃ×ｃ’個の動作）

【００６８】

しかし、代替の実施形態では、以下が使用され得る。

Ｃｏｎｖ２ｄ　１／１
（１×１×ｃ×ｔ＊ｃ個のパラメータ）
（ｈ×ｗ×１×１×ｃ×ｔ＊ｃ個の動作）
ＤＷＣｏｎｖ　３／ｓ
（３×３×１×ｔ＊ｃ個のパラメータ）
（ｈ／ｓ×ｗ／ｓ×３×３×１×ｔ＊ｃ個の動作）
Ｃｏｎｖ２ｄ　１／１
（１×１×ｔ＊ｃ×ｃ’個のパラメータ）
（ｈ／ｓ×ｗ／ｓ×１×１×ｔ＊ｃ×ｃ’個の動作）

【００６９】

上の代替実施形態では、より少数の計算が要求され、パラメータは、より少数である。

【００７０】

より高解像度での特徴の学習およびより低解像度へのそれらの射影は、特徴学習手順に利益を与える。さらに、ボトルネックブロックでの深さ方向畳み込みの利用は、計算効率をもたらし、メモリフットプリントを大幅に削減する。

【００７１】

さらなる実施形態では、残差接続が、ボトルネックブロックに組み込まれ、ボトルネック残差ブロックを形成する。残差接続は、入力ボリュームおよび出力ボリュームが同一の空間次元および同一個数の特徴マップを有する場合に限って、ボトルネックブロック内で使用され得る。残差接続は、ボトルネックブロックの入力からその出力に接続された直線の層を表す。追加の接続層は、乗算層にまたがるより効率的な勾配伝搬を可能にし、ニューラルネットワークのトレーニングを改善する。

【００７２】

一実施形態では、ステップＳ３０７で物体分類システムに使用されるボトルネック残差

ブロックは、表２に示された構造を有する。第１の層は、特徴マップの個数をｔ倍だけ増加させることによって入力ボリュームの次元を拡張する標準的な２次元畳み込み層である。

【表２】

表２

入力	演算子	出力
$h \times w \times c$ $h \times w \times tc$ $\frac{h}{s} \times \frac{w}{s} \times tc$	Conv2d 1/1, f DWConv 3/s, f Conv2d 1/1, -	$h \times w \times tc$ $\frac{h}{s} \times \frac{w}{s} \times tc$ $\frac{h}{s} \times \frac{w}{s} \times c'$

【００７３】

ボトルネック残差ブロックの第２の層では、深さ方向畳み込みフィルタが、シーンから特徴を抽出するのに使用される。深さ方向畳み込みは、ストライドｓを有し、したがって、特徴マップの出力次元は、空間的にｓ倍だけ縮小される。ストライドが１である畳み込み層では、特徴マップの空間サイズは影響を受けない。空間サイズの縮小は、後続のネットワーク層内での計算の回数を効果的に減少させる。畳み込みは、各深さで別々に計算されるので、ＤＷｃｏｎｖは、標準的な畳み込みに対して大幅に改善する。空間サイズの縮小は、後続のネットワーク層内での計算（パラメータではない）の数を効果的に減少させる。

【００７４】

点方向畳み込みがそれに続く深さ方向畳み込みは、標準的な畳み込み層と比較して、物体検出の正確さを大幅には低下させない。この情報は、まずチャネル方向レベルで計算され、これに、チャネル方向情報を結合する、より安価な標準的な畳み込みが続く。畳み込みはカーネル１×１のみを使用するので、より高速であり、より少数のパラメータを必要とする。

【００７５】

最後に、ボトルネック残差ブロックでは、深さ方向の分離可能な畳み込み層によって生成された出力ボリュームが、２次元畳み込みの第２の層を使用して、その元々の低次元表現に戻って射影され得る。第１および第２の２次元畳み込みに使用される拡張係数ｔは、この２つの層に関して同一である。

【００７６】

一実施形態では、ステップＳ３０７のボトルネック残差ブロックには、ピラミッドプーリングモジュールＳ３０９が続く。ピラミッドプーリングモジュールは、画像をより微細なレベルからより粗なレベルへの区分に分割し、それらの中の局所特徴を集約する。空間ピラミッドプーリングは、テストのために任意のサイズの画像／ウィンドウから表現を生成することと、トレーニング中に、変化するサイズまたはスケールの画像を供給することとを可能にする。

【００７７】

ピラミッドプーリングモジュールが、図５に、より詳細に示されている。ここでは、それは、４つのカーネルからなり、各カーネルの解像度は、それぞれ３２×６４、１６×３２、８×１６、および４×８である。カーネルは、ステップＳ３０７の特徴マップ出力に対して均等に分布される。結果の４つの特徴マップは、単一の特徴マップ出力を生成するために、双線形（bilinearly）にアップサンプリングされ、一緒に加算される。特徴マップが加算されることに留意することが重要である。これは、メモリ内にすべての解像度を記憶する必要を回避する。

【００７８】

表３は、一実施形態に従って大域コンテキストを取り込むのに使用され得る層の詳細を

示す。

【表３】

表３

入力	演算子	拡張係数	出力チャネル	繰返し	ストライド
128 x 256 x 64	ボトルネック	6	64	3	2
64 x 128 x 64	ボトルネック	6	96	3	2
32 x 64 x 96	ボトルネック	6	128	3	1
32 x 64 x 128	PPM	-	128	-	-

【００７９】

これは、図４にも絵図的に示されている。「ダウンサンプリングを学習する」モジュールの後で、ネットワークは、２つの分岐に分岐する。第１のネットワーク分岐は、９つのボトルネック残差ブロック４０９、４１１、４１３、４１５、４１７、４１９、４２１、４２３、４２５、および４２７と、これに続くピラミッドプーリングモジュール４２７とを備える。第２の分岐４４７は、スキップ接続を表す。

【００８０】

当業者によって了解されるように、ボトルネック残差演算子は、実際には、表２を参照して上で説明したように複数の層を介して実施される。しかし、図４では、簡易的に、ボトルネック残差演算子が単一のエンティティとして図示されている。

【表４】

表４

入力	ブロック	t	c	n	s
1024 × 2048 × 3	Conv2D	-	32	1	2
512 × 1024 × 32	DSConv	-	48	1	2
256 × 512 × 48	DSConv	-	64	1	2
128 × 256 × 64	ボトルネック	6	64	3	2
64 × 128 × 64	ボトルネック	6	96	3	2
32 × 64 × 96	ボトルネック	6	128	3	1
32 × 64 × 128	PPM	-	128	-	-
32 × 64 × 128	FFM	-	128	-	-
128 × 256 × 128	DSConv	-	128	2	1
128 × 256 × 128	Conv2D	-	19	1	1

【００８１】

第１および第２のボトルネック残差層４０９および４１１は、６４個の入力チャネルと、６４個の出力チャネルと、ストライド１と、６の拡張係数（ｔ）とを有する。

【００８２】

第３のボトルネック残差層４１３は、６４個の入力チャネルと、６４個の出力チャネルと、ストライド２と、６の拡張係数とを有する。

【００８３】

第４および第５のボトルネック残差層４１５および４１７は、６４個の入力チャネルと、６４個の出力チャネルと、ストライド１と、６の拡張係数とを有する。

【００８４】

第６のボトルネック残差層４１９は、６４個の入力チャネルと、９６個の出力チャネルと、ストライド２と、６の拡張係数とを有する。

【００８５】
第７および第８のボトルネック残差層４２１および４２３は、９６個の入力チャネルと、９６個の出力チャネルと、ストライド１と、６の拡張係数とを有する。
【００８６】
第９のボトルネック残差層４２５は、９６個の入力チャネルと、１２８個の出力チャネルと、ストライド２と、６の拡張係数とを有する。
【００８７】
大域特徴抽出器の最後の層は、ピラミッドプーリング層４２７である。ピラミッドプーリング層４２７は、１２８個の入力チャネルと１２８個の出力チャネルとを有する。
【００８８】
一実施形態では、ダウンサンプリングを学習するモジュールは、主に、大域特徴抽出器によって抽出された大域コンテキストを洗練する処理を担う。より高い正確さおよびよりよい物体分類結果を達成するために、ダウンサンプリングを学習するモジュール内の深さ方向の分離可能なブロックは、直接に使用される。深さ方向の分離可能な畳み込みのボトルネック実施態様は、達成される分類正確さより動作の速度が重要である場合に、多数の層からなるネットワーク分岐を好むが、直接手法は、低下した動作速度を犠牲にして、物体分類のより高い精度をもたらす。しかし、ダウンサンプリングを学習することにおけるより少数のネットワーク層は、より多数の動作を補償する。
【００８９】
一実施形態では、図３のステップＳ３０９で、単一の特徴融合モジュールが、ネットワークの２つの分岐の出力ボリュームをマージするのに使用される。特徴マップをマージするプロセスは、システムメモリ内に特徴を保持することを含む。単一の特徴融合ユニットは、低解像度デバイス要件に従う、より効率的な設計を考慮に入れたものである。
【００９０】
一実施形態では、２つの分岐の出力ボリュームは、特徴マップの数と空間次元との両方において異なる。より低解像度の分岐の特徴マップの空間次元は、フル解像度の分岐の特徴マップの空間次元より小さい。したがって、より低解像度のサブネットワーク分岐の出力は、アップサンプリング層４２９によって処理され、特徴マップは、４倍でアップスケーリングされる。特徴融合モジュールのアーキテクチャは、表５にも表されている。
【表５】
表５

分岐 -1	分岐 -4
- - Conv2d 1/1 -	*アップサンプリング x 4* *DWConv (膨張 4)* *3/1, f* *Conv2d 1/1, -*
Add, f	

【００９１】
さらなる実施形態では、アップサンプリング層に、１とは異なる膨張係数を有する深さ方向畳み込み層４３１が続く。膨張畳み込み層は、特徴マップ上の物体の間の空間を増大させる。膨張させる深さ方向畳み込みは、カーネルのサイズだけに影響し、具体的には、カーネルのサイズが、指数関数的に増大される。たとえば、１の膨張係数を有する深さ方向畳み込みは、３×３の元々のカーネルサイズを有するが、２の膨張係数を有する畳み込みは、７×７のカーネルサイズを有し、４の膨張係数を有する畳み込みは、１５×１５のカーネルサイズを有する。膨張が、カーネルサイズ７×７のものであるが、これが９つの計算だけを有することに留意されたい。
【００９２】

出力ボリュームの合計は、ボリュームが同一の空間次元と同一個数の特徴マップとを有する場合に限って実施され得る。したがって、２次元畳み込みの層が、２つの分岐の出力、それぞれ４３３および４３５が一緒に加算される前に、これらを一般化するのに使用される。この２つの畳み込み層は、２つの分岐の特徴マップが同一の次元を有することを保証する。

【００９３】

上で説明された実施形態では、畳み込み層４３３は、１２８個の入力チャネルおよび１２８個の出力チャネルを有するが、畳み込み層４３５は、４８個の入力チャネルおよび１２８個の出力チャネルを有する。

【００９４】

最終ステップでは、特徴マップは、４３７で単に直接に一緒に加算される。したがって、メモリ内に記憶される必要があるパラメータの個数は、増加しない。

【００９５】

一実施形態では、特徴融合ユニットには、分類器モジュール、図３のステップＳ３１１が続く。分類器のアーキテクチャは、表４と図４のブロック４３９、４４１、４４３、４４５、および４４７に関して説明され得る。分類器は、２つの深さ方向の分離可能な畳み込み演算４３９および４４１と点方向畳み込み４４３とを使用する。点方向畳み込み層４４３は、使用される都市景観セグメンテーションデータセット内の１９個のクラスのうち１つごとに１つの出力チャネルの、１９個の出力チャネルのみを備える。分類器の最後の２つの層は、画像の当初の空間次元を復元するアップサンプリング層４４５と、クラスラベルごとに確率スコアを作るｓｏｆｔ－ｍａｘ畳み込み層４４７である。特徴融合モジュールの後に配置される分類器の追加の畳み込み層は、物体分類のシステムの正確さを改善する。

【００９６】

一実施形態による分類器モジュール内の畳み込み層のパラメータが、表５に提示されている。

【表６】

表5

入力	演算子	出力チャネル	繰返し	ストライド
128 x 256 x 128	DWConv	128	1	1
128 x 256 x 128	Conv2d (1x1)	128	1	1
128 x 256 x 128	DWConv	128	1	1
128 x 256 x 128	Conv2d(1x1)	128	1	1
128 x 256 x 128	Conv2d (1x1)	19	1	1
128 x 256 x 19	アップサンプリング	19	1	1
1048 x 2048　x 19	Soft-max	19	1	1

【００９７】

図６は、入力画像内に示された物体にラベルを付けることのできる、物体分類ニューラルネットワークをトレーニングする例のプロセスのフロー図を示す。ニューラルネットワークは、トレーニングの多数のサンプルを処理することと、すべてのサンプルについて、ニューラルネットワークによって生成された出力とトレーニングサンプル内で指定されるターゲット出力との間の誤差に従って各パラメータの重みを調整することとによってトレーニングされ得る。トレーニングされた後に、ニューラルネットワークは、システム、たとえば図２のニューラルネットワークシステム内で展開され得る。トレーニング手順は、１つまたは複数のコンピュータによって実行され得る。

【００９８】

ステップＳ６０１では、トレーニングシステムが、トレーニングデータのセットを入手する。各データセットは、トレーニング画像とターゲット出力とを備える。入力画像は、１つまたは複数の物体を示す画像の表現である。たとえば、画像は、自律車両または半自律車両の付近に配置された物体を含むことができる。トレーニングデータセットによって表されるトレーニング画像は、互いとは異なり、同様の物体を含んでも含まなくてもよい。トレーニングデータセットは、ニューラルネットワークをトレーニングするのに使用され得る、有限個数のシーンを備える。一実施形態では、標準的なデータ増補技法が、数サンプル画像を拡張するのに使用され得る。増補技法は、０．５から２までの範囲内のランダムなスケール係数と、水平フリップと、変更された色相と、変更された飽和度と、変更された輝度と、変更されたコントラストとを含むが、これに限定はされない。

【００９９】

トレーニングデータセットのトレーニングターゲット出力は、ニューラルネットワークによって生成されるべき物体分類ネットワークの所望の出力を表す。ターゲット出力は、ニューラルネットワークの実際の出力と比較され、重み付けパラメータは、ターゲット出力と生成された出力との間の誤差が縮小されるようにするために調整される。ステップＳ６０３では、ニューラルネットワークが、内部パラメータの現在値を使用してサンプル入力画像を処理し、出力画像を生成する。

【０１００】

ステップＳ６０５では、ニューラルネットワークの予測された出力が、トレーニングデータセットのターゲット出力と比較され、予測の誤差が推定される。

【０１０１】

その結果、ステップＳ６０７では、各内部パラメータの重みが、予測された出力とターゲット出力との間の誤差が最小値まで減らされるようにするために調整される。

【０１０２】

ステップ６０９では、ニューラルネットワークが、トレーニングデータの異なるセットを与えられ、トレーニングは、トレーニング手順を繰り返し、予測された出力とターゲット出力とのより小さい誤差が達成されるようにするために、ニューラルネットワークの内部パラメータを調整するために、ステップＳ６０３に戻る。

【０１０３】

一実施形態では、モデルは、交差エントロピー損失を使用してトレーニングされ、トレーニング中に、重み付き補助損失（weighted auxiliary loss）が、ダウンサンプリングを学習するモジュールおよび大域特徴抽出モジュールの終りで使用される。この形での損失の重み付けは、セマンティックセグメンテーションの意味のある特徴がダウンサンプリングを学習するモジュールおよび大域特徴抽出モジュールによって抽出され、ネットワークの他の副部分とは独立に学習されることを保証する。一実施形態では、補助損失の重みに０．４がセットされた。一実施形態では、これが、４０７および４２７の後に追加の出力を生成することによって達成される（すなわち、新しいｓｏｆｔｍａｘ層が、このステージに導入され、ネットワークの分岐からフォークする）。ｓｏｆｔｍａｘ層の出力が評価される（タスクはセグメンテーションでもある）。３つの層の出力が、重みを更新するのに使用される。勾配降下法が使用され、これが０と１との間の確率値を与えるので、ｓｏｆｔｍａｘが、トレーニング中に使用される。推論中に、ｓｏｆｔｍａｘとａｒｇｍａｘとの両方の関数が単調に増加するので、高コストのｓｏｆｔｍａｘ計算が、ａｒｇｍａｘに置換される。ａｒｇｍａｘは、１または０、すなわち、物体が存在するまたは存在しない、を用いてデータにラベルを付ける。

【０１０４】

ニューラルネットワークのトレーニングプロセスの一実施形態では、ネットワーク枝刈りが実施される。トレーニングの最初のステージでは、特徴マップの個数が２倍にされ、トレーニングは、上で説明された手順を使用して行われる。パラメータの個数は、元々のサイズの１．７５倍、１．５倍、１．２５倍、および１倍に徐々に減らされ、ここで、トレーニング手順は、パラメータのそれぞれの減少の後に繰り返される。ネットワーク枝刈

りは、同一の性能を保ちながらネットワーク内のパラメータの個数を効果的に削減する。これは、ニューラルネットワークアーキテクチャ内で使用されていない特徴を除去することによって達成される。さらに、ネットワーク枝刈りは、必要な特徴だけがネットワークによって学習されるので、より効率的な学習を可能にする。

【０１０５】

したがって、枝刈りは、通常、パラメータを減らすのに使用される。しかし、本明細書で説明される実施形態では、パラメータの個数が許容できるものなので、枝刈りは、性能を高めるために実行される。したがって、ネットワークの表現力は、２倍にされる（パラメータの個数を２倍にする）。今や、ネットワークは、より多数のフィルタが存在する（個数を２倍にする）ので、はるかにより低速である。しかし、枝刈りは、フィルタの元の個数にもう一度達するのに使用される。ここで、枝刈りは、フィルタの個数を拡大し（より多数を可能にし）、その後、良好である１回を選択する。

【０１０６】

したがって、ニューラルネットワークのトレーニングプロセスのこの実施形態では、フィルタの個数は、トレーニング手順の初めに２倍にされる。これがどのように行われるのかは、図７を参照して説明される。

【０１０７】

ステップＳ７０１では、層ごとのフィルタの個数が２倍にされる。ステップＳ７０３では、フィルタのチャネル数が、前の層の出力のチャネル数と一致させられる。たとえば、各層が、入力サイズｈ×ｗと出力サイズｈ’×ｗ’を有するものとする。すべてのこれらの層が、使用されるフィルタの個数に関する深さ／チャネルを有する。すなわち、ｈ×ｗ×ｃを与えられて、サイズ３×３（たとえば）のｃ’個のフィルタが、ｈ’×ｗ’×ｃ’を作るのに使用され得る。ステップＳ７０１と同様に、前の層のフィルタの個数が２倍にされる場合に、ｈ×ｗ×２ｃ個の入力があり、サイズ３×３×２ｃの２ｃ’個のフィルタが使用される。

【０１０８】

ステップＳ７０５では、ネットワークが、図６に関して説明されたようにトレーニングされる。ステップＳ７０７では、容量に達したかどうかすなわち、フィルタのターゲット数に達したかどうかが決定される。

【０１０９】

ステップＳ７０９では、第１の層内の最弱のフィルタ（画像入力を有する）が、識別され、除去される。これらのフィルタは、多数の異なる行列を識別する可能性があり、一実施形態では、ｌ１和が使用される。たとえば、上の次元を使用すると、ｈ×ｗ×１．５ｃが、今は入力である。次の層のフィルタは、サイズ３×３×２ｃであり、したがって、ステップＳ７１１では、前の層の除去されるフィルタに関係する重みが、サイズ３×３×１．５ｃのフィルタを得るために除去される。その後、ステップＳ８１３では、現在の層内の最弱のフィルタが、決定され、このプロセスが継続される。

【０１１０】

ステップＳ７１５では、ステップＳ７１１およびＳ７１３でのフィルタの除去によって影響を受ける層がまだあるかどうかを調べるためにチェックされる。ある場合には、このプロセスはステップＳ７１１にループバックする。最弱のフィルタおよび重みのすべてが除去された後に、このプロセスは、トレーニングステップＳ７０７に戻る。ネットワークがトレーニングされた後に、フィルタの個数がさらに減らされる（たとえば、ここでは、フィルタの個数が、２×から１．７５×に、１．５×に、１．２５×に、１×に減らされる）べきかについてチェックされる。さらなる削減が要求される場合には、このプロセスは、さらなるフィルタを除去するためにステップＳ７０９に移る。そうではない場合には、このプロセスは終了し、ネットワークは、トレーニングされ、枝刈りされる。

【０１１１】

一実施形態では、バッチ正規化が、トレーニング中にダウンサンプリングを学習するモジュールのすべての層の後で使用される。バッチ正規化は、ニューラルネットワークの各

層が他の層とはより独立に学習することを可能にする。バッチ正規化は、前の層の特徴マップ内のアクティブ化をスケーリングする。ニューラルネットワーク全体のすべてのアクティブ化が、所与の範囲内なので、大きすぎる値または小さすぎる値に関連するアクティブ化はない。これは、より高い学習速度と改善された特徴学習とを可能にする。

【０１１２】

一実施形態では、ニューラルネットワークノードのセットの脱落が、ｓｏｆｔ－ｍａｘ層の前に実施される。脱落は、ネットワーク内のニューロンの間の相互に依存する学習を減らすトレーニング手法である。トレーニング中に、ノードのランダムなセットが、ネットワークから脱落され、その結果、ネットワークの縮小された版が作成されるようになる。ネットワークの縮小された版は、ニューラルネットワークの他のセクションとは独立に学習し、したがって、ニューロンが互いの間の共依存関係を展開するのを防ぐ。

【０１１３】

上を実証するために、上の切除研究が、都市景観データセットを使用して行われ［M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, および B. Schiele、The Cityscapes dataset for semantic urban scene understanding. ＣＶＰＲ、２０１６年］、都市景観テストセット、すなわち都市景観ベンチマークサーバに関する性能を報告する。

【０１１４】

実験は、ＣＵＤＡ　９．０およびｃｕＤＮＮ　Ｖ７を用い、Ｎｖｉｄｉａ　Ｔｉｔａｎ　Ｘ（Ｍａｘｗｅｌｌ、３０７２個のＣＵＤＡコア）またはＮｖｉｄｉａＴｉｔａｎ　Ｘｐ（Ｐａｓｃａｌ、３８４０個のＣＵＤＡコア）を有するワークステーション上で実行された。ＲｅＬＵが、ＲｅＬＵ６と比較して、達成される、より高速のトレーニングおよびよりよい正確さとに起因して、非線形性関数として使用された。トレーニング中に、バッチ正規化が、すべての層で使用され、脱落が、ｓｏｆｔ－ｍａｘ層の前のみで使用される。推論中に、バッチ正規化のパラメータは、親層の重みおよびバイアスとマージされる。深さ方向畳み込み層では、我々は、ｌ２正則化が不要であることを見出した。一実施形態では、他の層に関して、ｌ２正則化は０．０００００４である。

【０１１５】

ラベル付けされたトレーニングデータが制限されたので、ランダムスケール０．５から２までと、水平フリップと、変更された色相と、変更された飽和度と、変更された輝度と、変更されたコントラストという標準的なデータ増補技法が、すべての実験で適用された。

【０１１６】

Ｆａｓｔ－ＳＣＮＮのモデルは、Ｐｙｔｈｏｎを使用するＴｅｎｓｏｒＦｌｏｗ機械学習プラットフォームを用いてトレーニングされる。０．９の運動量およびバッチサイズ１２を有する確率的勾配降下法（ＳＧＤ）が使用される。さらに、ポリ学習レート（poly learning rate）が、ベースレート０．０４５および電力０．９８を用いて適用された。エポックの最大個数は、事前トレーニングが使用されないので１０００にセットされる。

【０１１７】

都市景観は、ドイツ内の５０の異なる都市からの街路シーン内の画像の多様なセットを含む、セマンティックセグメンテーションの大規模データセットである。合計して、それは、２５０００枚の注釈付き１０２４×２０４８画素画像であり、そのうちの５０００枚は、高い画素正確さでのラベルを有し、２００００枚は、弱く注釈付けされている。本明細書で提示される実験では、５０００枚の画像だけすなわち、都市景観評価サーバ上で評価され得る２９７５枚の画像のトレーニングセットと、５００枚の画像の妥当性検査セットと、１５２５枚の試験画像とが、高いラベル品質を伴って使用された。

【０１１８】

一実施形態では、ＩｍａｇｅＮｅｔを用いる事前トレーニングが使用された。

【０１１９】

都市景観は、３０個のクラスラベルをも付与するが、１９個のクラスだけが、評価に使

用される。結果は、平均インターセクションオーバユニオン（ｍｅａｎ　ｉｎｔｅｒｓｅｃｔｉｏｎ－ｏｖｅｒ－ｕｎｉｏｎ）（ｍＩｏＵ）として報告され、ランタイム評価は、転送推論時間（ｆｏｒｗａｒｄ　ｉｎｆｅｒｅｎｃｅ　ｔｉｍｅ）を測定するためにシングルスレッド式ＣＰＵおよびＧＰＵ内で実行される。バーンインのために、１００フレームが使用され、フレーム毎秒（ｆｐｓ）測定のために、１００フレームの平均値が報告される。

【０１２０】

ｆａｓｔ－ＳＣＮＮの全体的な性能は、都市景観の差し控えられたテストセットに対して評価される。表６には、Ｎｖｉｄｉａ　Ｔｉｔａｎ　Ｘ（Ｍａｘｗｅｌｌ、３０７２個のＣＵＤＡ）と、「*」によって表されるＮｖｉｄｉａ　Ｔｉｔａｎ　Ｘｐ（Ｐａｓｃａｌ、３８４０個のＣＵＤＡコア）との両方に関して、異なる解像度でのｆｐｓ単位で比較されたランタイムがある。ｆａｓｔ－ＳＣＮＮの２つの版すなわち、ｓｏｆｔ－ｍａｘ出力（我々のｐｒｏｂ）および物体ラベル出力（我々のｃｌｓ）が示されている。

【表７】

表６

	1024 x 2048	512 x 1024	256 x 512
我々の prob	47.5	149.4	310.7
我々の cls	57.8	171.6	379.7
我々の prob*	78.5	244.3	475.0
我々の cls*	91.0	273.5	531.4

【０１２１】

都市景観テストセットを使用するＦａｓｔ－ＳＣＮＮのクラスおよびカテゴリｍＩｏＵが、表７に提示されている。Ｆａｓｔ－ＳＣＮＮは、６８．０％　ｍＩｏＵを達成する。このモデルは、少メモリ組み込みデバイス用に設計され、１．１百万パラメータだけを使用する。Ｆａｓｔ－ＳＣＮＮの結果は、定量分析のために図９に表示されている。第１の列には、入力ＲＧＢ画像があり、第２の列は、グラウンドトゥルースラベルであり、第３の列は、Ｆａｓｔ－ＳＣＮＮの出力である。

【表８】

表７

モデル	クラス	カテゴリ	パラメータ
Fast-SCNN	68.0	84.7	01.11

【表９】

表８

入力サイズ	クラス	FPS
1024 x 2048	68.0	91.0
512 x 1024	62.8	273.5
256 x 512	51.9	531.4

【０１２２】

Ｆａｓｔ－ＳＣＮＮが、少ない容量を有するように特に設計されているので、少メモリであることの理由は、組み込みデバイス上での実行を可能にし、よりよい一般化が期待される。提案されるネットワークの性能は、事前トレーニングの有無を伴い、追加の弱くラ

ベル付けされたデータの有無に関連して評価された。結果は、表９に提示されている。事前トレーニングに関して、ＩｍａｇｅＮｅｔデータベースが使用され、特徴融合モデルは、平均プーリングによって置換され、分類モジュールは、ｓｏｆｔｍａｘ層だけを備える。ＩｍａｇｅＮｅｔに対する事前トレーニングは、しばしば、正確さと一般性とを押し上げる。ＩｍａｇｅＮｅｔ事前トレーニングを用いるＦａｓｔ－ＳＣＮＮの正確さは、都市景観の妥当性検査セットに関して６９．１５％　ｍＩｏＵであるが、Ｆａｓｔ－ＳＣＮＮは、事前トレーニングなしで６８．６２％　ｍＩｏＵを達成する。

【０１２３】

さらに、都市景観の都市道路とＩｍａｇｅＮｅｔの分類タスクとの間のオーバーラップが制限されるので、Ｆａｓｔ－ＳＣＮＮが、両方の領域の制限された能力に起因して利益を得ない可能性があると仮定することは、穏当である。したがって、都市景観によって付与される追加の２００００個の粗にラベル付けされた画像が、類似する領域からのものなので、組み込まれた。それでも、粗なトレーニングデータ（ＩｍａｇｅＮｅｔありまたはＩｍａｇｅＮｅｔなし）を用いてトレーニングされたＦａｓｔ－ＳＣＮＮは、互いと同様に、事前トレーニングなしの元々のＦａｓｔ－ＳＣＮＮに対するわずかな改善のみを伴って実行する。

【０１２４】

低容量Ｆａｓｔ－ＳＣＮＮが、Ｉｍａｇｅｎｅｔを用いる事前トレーニングから大きくは利益を得ないと結論することができる。同様の結果が、積極的なデータ増補およびより多数のエポックを使用することによって達成され得る。

【表１０】

表9

モデル	クラス
Fast-SCNN	68.62
Fast-SCNN + ImageNet	69.15
Fast-SCNN + 粗	69.22
Fast-SCNN + 粗 + ImageNET	69.19

【０１２５】

図９に、トレーニング曲線を示す。粗データを用いるＦａｓｔ－ＳＣＮＮは、弱いラベル品質のゆえに反復に関して低速でトレーニングする。ＩｍａｇｅＮｅｔ事前トレーニング版の両方は、早期エポック（トレーニングセットのみに関して４００エポックまで、追加の粗にラベル付けされたデータを用いてトレーニングされる時に１００エポックまで）に関してよりよく実行する。これは、Ｆａｓｔ－ＳＣＮＮが一からトレーニングされる時に、同様の正確さに達するためにより長くトレーニングされる必要があることを意味する。

【０１２６】

上の実施形態は、高解像度画像（１０２４×２０４８画素）に関するリアルタイムより高速の物体分類（９１．０ｆｐｓ）のためのネットワークに関する。多重分岐ネットワークの計算コストの共有は、ランタイム効率をもたらす。上のアーキテクチャでは、スキップ接続が、空間詳細の回復に関して有益であることを示す。スキップ接続は、小さいサイズの境界および物体の周囲で特に有益である、図１０。

【０１２７】

さらに、上の研究は、十分に長くトレーニングされた場合に、追加の補助タスクに対するモデルの大規模事前トレーニングが、低容量ディープ畳み込みニューラルネットワークに関して必要ではないことを示す。

【０１２８】

図１１は、実施形態に従って方法を実施するのに使用され得るハードウェアの概略図で

ある。これが、一例にすぎず、他の配置が使用され得ることに留意されたい。

【０１２９】

ハードウェアは、計算セクション９００を備える。この特定の例では、このセクションの構成要素は、一緒に説明される。しかし、これらが、必ずしも同一位置に配置されないことを了解されたい。

【０１３０】

コンピューティングシステム９００の構成要素は、処理ユニット９１３（中央処理装置、ＣＰＵなど）と、システムメモリ９０１と、システムメモリ９０１を含む様々なシステム構成要素を処理ユニット９１３に結合するシステムバス９１１とを含むがこれに限定されない。システムバス９１１は、メモリバスもしくはメモリコントローラと、周辺バスと、様々なバスアーキテクチャのいずれかを使用するローカルバスなどとを含む複数のタイプのバス構造のいずれとすることもできる。計算セクション９００は、バス９１１に接続された外部メモリ９１５をも含む。

【０１３１】

システムメモリ９０１は、読取専用メモリなど、揮発性メモリ／または不揮発性メモリの形のコンピュータ記憶媒体を含む。スタートアップ中などにコンピュータ内の要素の間での情報の転送を助けるルーチンを含む基本入出力システム（ＢＩＯＳ）９０３が、通常はシステムメモリ９０１内に記憶される。さらに、システムメモリは、ＣＰＵ　９１３によって使用中のオペレーティングシステム９０５と、アプリケーションプログラム９０７と、プログラムデータ９０９とを含む。

【０１３２】

また、インターフェース９２５が、バス９１１に接続される。インターフェースは、コンピュータシステムがさらなるデバイスから情報を受信するためのネットワークインターフェースとすることができる。インターフェースは、ユーザがある種のコマンドなどに応答することを可能にするユーザインターフェースとすることもできる。

【０１３３】

この例では、ビデオインターフェース９１７が設けられる。ビデオインターフェース９１７は、グラフィックス処理メモリ９２１に接続されたグラフィックス処理ユニット９１９を備える。

【０１３４】

グラフィックス処理ユニット（ＧＰＵ）９１９は、この多重並列呼出しの動作に起因して、上で説明される方法に特によく適する。したがって、一実施形態では、処理は、ＣＰＵ９１３とＧＰＵ９１９との間で分割され得る。

【０１３５】

一実施形態では、ＧＰＵは、低電力ＧＰＵチップであるＮＶＩＤＩＡ　Ｊｅｔｓｏｎ　ＴＸ２である。

【０１３６】

一実施形態では、専用コンピューティングデバイス９００は、各カメラ（図１参照）へのリンクを設けられる。上で図２から図４までに関して説明されるアーキテクチャは、浮動小数点計算の必要を回避し、したがって、コンピューティングデバイスは、車両上のカメラと結合されるなど、低電力位置によく適する。

【０１３７】

上で説明されるアーキテクチャは、ＧＰＵを使用する携帯電話機にも特に役立つ。

【０１３８】

ある種の実施形態が説明されたが、これらの実施形態は、例としてのみ提示され、本発明の範囲を限定することは意図されていない。実際に、本明細書で説明される新規のデバイスおよび方法は、様々な他の形で具現化され得、さらに、本明細書で説明されるデバイス、方法、および製品の形態における様々な省略、置換、および変更が、本発明の趣旨から逸脱せずに行われ得る。添付の特許請求の範囲およびその均等物は、本発明の範囲および趣旨に含まれ得るものとしてそのような形態または修正を包含することが意図されてい

る。

【図１】

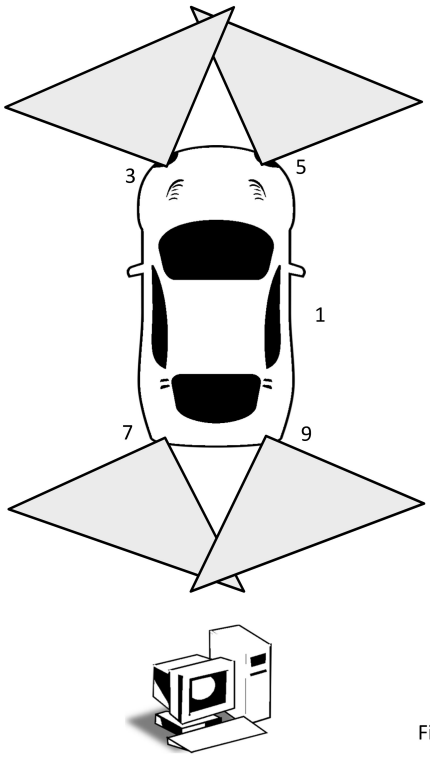

Figure 1

【図２】

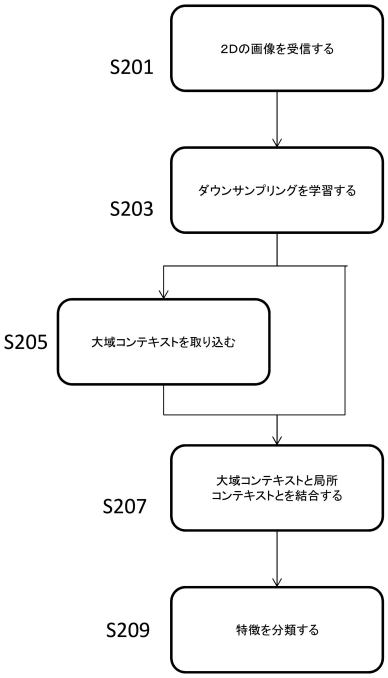

Figure 2

【図３】

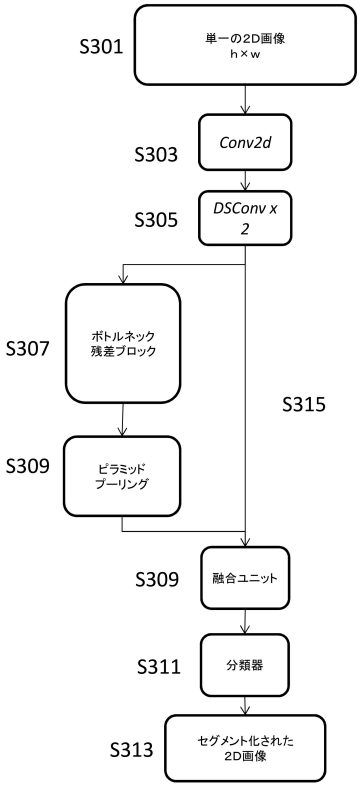

Figure 3

【図４】

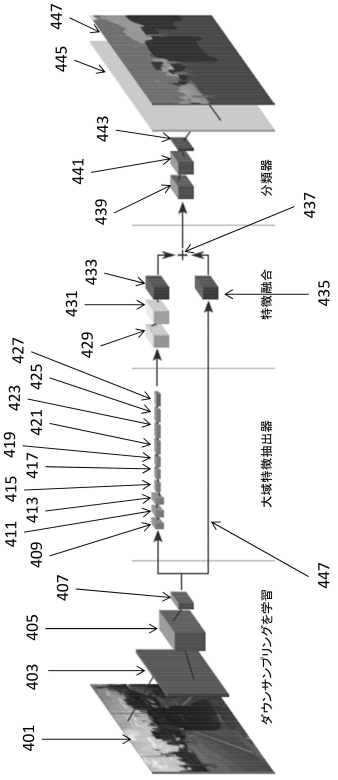

Figure 4

【図５】

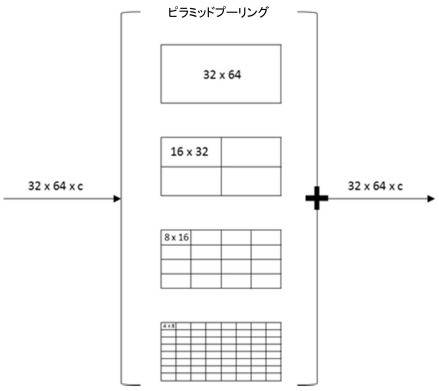

Figure 5

【図６】

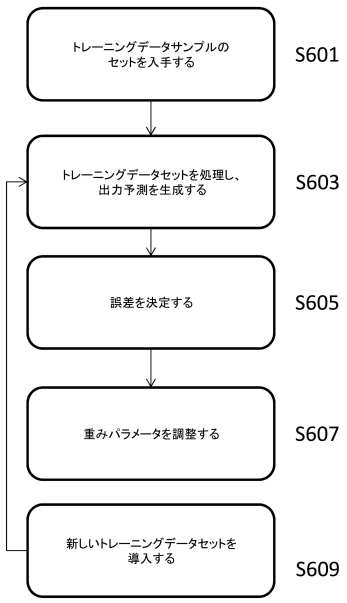

Figure 6

【図７】

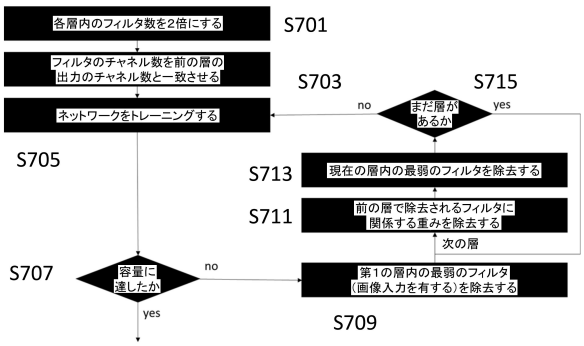

Figure 7

【図８】

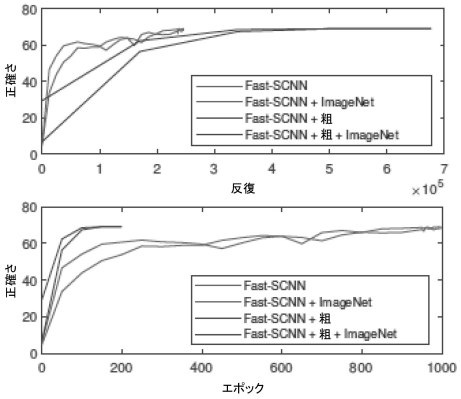

Figure 8

【図９】

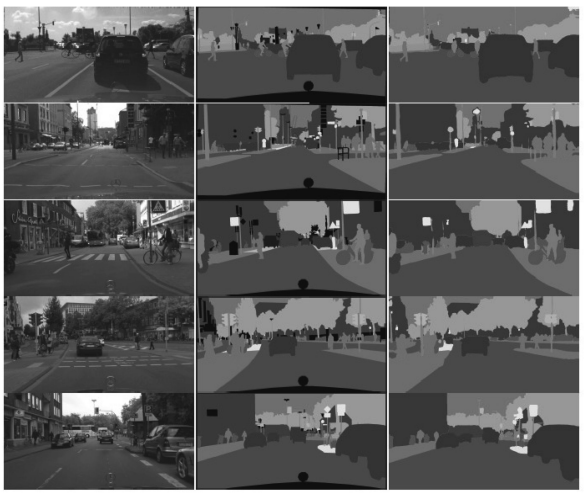

Figure 9

【図１０】

Figure 10

【図１１】

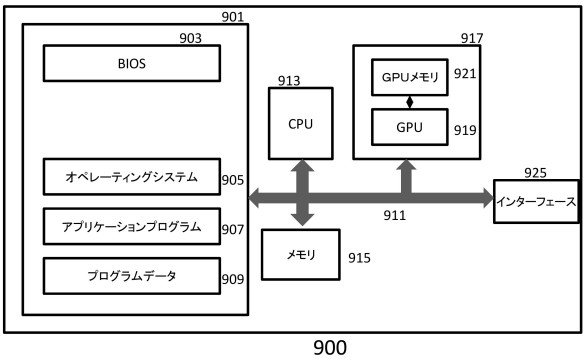

Figure 11

フロントページの続き

(72)発明者　ステファン　リウィッキ
イギリス国、　シービー４・０ジーゼット、　ケンブリッジシャー、　ケンブリッジ、　ミルトン・ロード、　ケンブリッジ・サイエンス・パーク　２０８、　トーシバ・リサーチ・ヨーロッパ・リミテッド、ケンブリッジ・リサーチ・ラボラトリー内
(72)発明者　ロベルト　シポラ
イギリス国、　シービー４・０ジーゼット、　ケンブリッジシャー、　ケンブリッジ、　ミルトン・ロード、　ケンブリッジ・サイエンス・パーク　２０８、　トーシバ・リサーチ・ヨーロッパ・リミテッド、ケンブリッジ・リサーチ・ラボラトリー内
(72)発明者　ルドラ　プラサド　パウデル　カルマタ
イギリス国、　シービー４・０ジーゼット、　ケンブリッジシャー、　ケンブリッジ、　ミルトン・ロード、　ケンブリッジ・サイエンス・パーク　２０８、　トーシバ・リサーチ・ヨーロッパ・リミテッド、ケンブリッジ・リサーチ・ラボラトリー内

審査官　岡本　俊威

(56)参考文献　国際公開第２０１８／１４５０２８（WO，Ａ１）
特開２０２０－０７１８６２（ＪＰ，Ａ）

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