(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7177062号
(P7177062)

(45)発行日　令和4年11月22日(2022.11.22)

(24)登録日　令和4年11月14日(2022.11.14)

(51)国際特許分類
Ｇ０６Ｔ　7/593(2017.01)

ＦＩ
Ｇ０６Ｔ　7/593

請求項の数　20　（全19頁）

(21)出願番号　特願2019-535986(P2019-535986)
(86)(22)出願日　平成29年9月12日(2017.9.12)
(65)公表番号　特表2019-526878(P2019-526878A)
(43)公表日　令和1年9月19日(2019.9.19)
(86)国際出願番号　PCT/GB2017/052671
(87)国際公開番号　WO2018/046964
(87)国際公開日　平成30年3月15日(2018.3.15)
審査請求日　令和2年9月9日(2020.9.9)
(31)優先権主張番号　1615470.0
(32)優先日　平成28年9月12日(2016.9.12)
(33)優先権主張国・地域又は機関
英国(GB)

(73)特許権者　519087723
ナイアンティック， インコーポレイテッド
ＮＩＡＮＴＩＣ，ＩＮＣ．
アメリカ合衆国 カリフォルニア州 ９４１１１ サンフランシスコ ワンフェリービルディング スィート ２００
One Ferry Building, Suite 200 San Francisco, CA 94111 United States of America
(74)代理人　100110928
弁理士　速水 進治
(74)代理人　100127236
弁理士　天城 聡

最終頁に続く

(54)【発明の名称】　統計モデルを用いた画像データからの深度予測

(57)【特許請求の範囲】
【請求項１】
コンピュータによって実現される方法であって、
単一入力カラー画像から視差値を予測するためのモデルであって、左画像と右画像とを含む少なくとも１つの入力された両眼ステレオ画像ペアを用いて、
前記右画像および前記左画像のいずれか一方を用いて、前記左画像に適用されたときに予測される右画像の再構築を可能にする左から右への視差値と、前記右画像に適用されたときに予測される左画像の再構築を可能にする右から左への視差値と、を予測すること、及び
前記左から右への予測視差値と前記右から左への予測視差値との間の整合性を高めるコスト関数に基づいて前記モデルを更新すること、
によって訓練された前記モデルを提供するステップと、
前記モデルを用いて前記単一入力カラー画像から前記視差値を生成するステップと、
前記単一入力カラー画像から生成される前記視差値を用いて、前記単一入力カラー画像に対応する推定深度データを算出するステップと、を含む方法。
【請求項２】
前記モデルは、前記左から右への予測視差値を用いて、右から左への投影視差値を計算すること、および、
前記右から左への予測視差値を用いて、左から右への投影視差値を計算すること、によってさらに訓練される、請求項１に記載の方法。

【請求項３】
前記右から左への投影視差値は、前記左から右への予測視差値をサンプリングし、かつ前記サンプリングされたデータに前記右から左への予測視差値を適用することで計算され、
前記左から右への投影視差値は、前記右から左への予測視差値をサンプリングし、かつ前記サンプリングされたデータに前記左から右への予測視差値を適用することで計算される、請求項２に記載の方法。
【請求項４】
前記コスト関数は、前記左から右および右から左への予測視差値と前記左から右および右から左への投影視差値との間の整合性を高めるための視差整合性構成要素を含む、請求項２に記載の方法。
【請求項５】
前記モデルは、前記左画像のサンプリングされたピクセルをずらすために前記左から右への予測視差値を適用することで、前記ステレオペアの前記右画像を再構築すること、および、
前記右画像のサンプリングされたピクセルをずらすために前記右から左への予測視差値を適用することで、前記ステレオペアの前記左画像を再構築すること、によってさらに訓練される、請求項１に記載の方法。
【請求項６】
前記サンプリングは、バイリニア補間を含む、請求項３に記載の方法。
【請求項７】
前記コスト関数は、再構築された前記予測される左画像及び前記予測される右画像と、前記ステレオペアの前記左画像および前記右画像との間の画像再構築誤差を最小にするための再構築アピアランスマッチング構成要素をさらに含む、請求項５に記載の方法。
【請求項８】
前記コスト関数は、前記左から右および右から左への予測視差値を局所的に平滑化するための平滑化構成要素をさらに含む、請求項７に記載の方法。
【請求項９】
前記コスト関数は、前記左から右および右から左への予測視差値と前記左から右および右から左への投影視差値との間の整合性を高めるための視差整合性構成要素、前記平滑化構成要素、及び前記再構築アピアランスマッチング構成要素の重み付き和を実現する、請求項８に記載の方法。
【請求項１０】
前記モデルは、各処理ノードが少なくとも一つの重み値を有する処理ノードの構造化された配置を含む畳み込みニューラルネットワーク（convolutional neural network: CNN）を含む、請求項１に記載の方法。
【請求項１１】
前記畳み込みニューラルネットワークは、前記コスト関数の逆伝播構成要素により訓練される、請求項１０に記載の方法。
【請求項１２】
前記モデルは、
前記入力された両眼ステレオ画像ペアの前記左画像および前記右画像を複数の空間解像度でアップサンプリング及びアップコンボリューションし、
左から右の視差値および右から左の視差値のそれぞれを各空間解像度で予測することによってさらに訓練され、
前記モデルは、前記左から右への予測視差値と前記右から左への予測視差値との間の整合性を各空間解像度で高めるコスト関数に基づいて更新される、請求項１に記載の方法。
【請求項１３】
前記コスト関数は、前記空間解像度に応じて前記左から右への予測視差値と前記右から左への予測視差値との間の整合性の重み付き強化を含む、請求項１２に記載の方法。
【請求項１４】

前記両眼ステレオ画像ペアは、既知のカメラ焦点長を有しかつ既知の基線距離だけ離れているそれぞれのカメラによって同時に撮像され、それによって前記左から右および右から左への予測視差値から対応する深度データが計算される、請求項１に記載の方法。
【請求項１５】
前記両眼ステレオ画像ペアは、修正されかつ時間的にアラインされたステレオペアである請求項１４に記載の方法。
【請求項１６】
デジタル画像は、前記画像を撮像した前記それぞれのカメラの属性を定義するメタデータで注釈付けされる、請求項１５に記載の方法。
【請求項１７】
コンピュータによって実現される方法であって、
単一入力カラー画像から視差値を予測するためのモデルを定義するデータを記憶するステップと、
左画像と右画像とを含む少なくとも１つの入力された両眼ステレオ画像ペアを用いた前記モデルの訓練を、
前記右画像および前記左画像のいずれか一方を用いて、前記左画像に適用されたときに予測される右画像の再構築を可能にする左から右への視差値、および、前記右画像に適用されたときに予測される左画像の再構築を可能にする右から左への視差値を予測すること、及び
前記左から右への予測視差値と前記右から左への予測視差値との間の整合性を高めるコスト関数に基づいて前記モデルを更新すること、
によって行うステップと、を含み、
前記訓練されたモデルは、
前記訓練されたモデルを用いて前記単一入力カラー画像から前記視差値を生成すること、および、
前記単一入力カラー画像から生成された前記視差値を用いて、前記単一入力カラー画像に対応する推定深度データを算出すること、
によって前記単一入力カラー画像から深度画像を生成するために使用される、方法。
【請求項１８】
前記単一入力カラー画像はカメラによって撮影される、請求項１７に記載の方法。
【請求項１９】
演算装置によって実行されると、
単一入力カラー画像から視差値を予測するためのモデルであって、左画像と右画像とを含む少なくとも１つの入力された両眼ステレオ画像ペアを用いて、
前記右画像および前記左画像のいずれか一方を用いて、前記左画像に適用されたときに予測される右画像の再構築を可能にする左から右への視差値と、前記右画像に適用されたときに予測される左画像の再構築を可能にする右から左への視差値と、を予測すること、及び
前記左から右への予測視差値と前記右から左への予測視差値との間の整合性を高めるコスト関数に基づいて前記モデルを更新すること、
によって訓練された前記モデルを提供するステップと、
前記モデルを用いて前記単一入力カラー画像から前記視差値を生成するステップと、
前記単一入力カラー画像から生成される前記視差値を用いて、前記単一入力カラー画像に対応する推定深度データを算出するステップと、
を含む動作を前記演算装置に行わせる命令を記憶する非一時的なコンピュータ可読媒体。
【請求項２０】
前記動作は、
前記入力された両眼ステレオ画像ペアの前記左画像および前記右画像を複数の空間解像度でアップサンプリング及びアップコンボリューションし、
左から右の視差値および右から左の視差値のそれぞれを各空間解像度で予測すること、

を更に含み、
前記モデルは、前記左から右への予測視差値と前記右から左への予測視差値との間の整合性を各空間解像度で高めるコスト関数に基づいて更新される、請求項１９に記載の非一時的なコンピュータ可読媒体。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、全体として、画像データ処理システムに関し、より具体的には、訓練された統計モデルを用いた画像データからの深度データの予測に関する。
【背景技術】
【０００２】
画像からの奥行き推定は、コンピュータビジョン分野において長い歴史を有する。SFM（structure from motion）、shape from X、両眼式、及びMVS（multi-view stereo）に依拠して有益なアプローチがあった。しかし、これらの技術のほとんどは、注目シーンの複数の観察が可能であるという仮定に依存している。これらの観察は、複数の視点、または異なる照明の条件下におけるシーンの観察といった形で得ることができる。この制限を克服するために、例えば、L. Ladicky, J. Shi 及び M. Pollefeys による「Pulling Things Out Of Perspective」（CVPR 2014）と、D. Eigen, C. Puhrsch 及び R. Fergusによる「Depth Map Prediction From A Single Image Using A Multi-Scale Deep Network」（NIPS 2014）と、F. Liu, C. Shen, G. Lin 及び I. Reid による「Learning Depth From Single Monocular Images Using Deep Convolutional Neural Fields」（PAMI 2015）などにおいて論議されているように、最近、教師あり学習の問題として、単一の入力画像しかない単眼深度推定の課題を提示する研究の数が急増している。しかしながら、このような研究に記述された方法では、オフラインで大量のグラウンドトゥルース深度データを用いてモデルが訓練され、そのモデルを用いて画像内の各ピクセルの深度を直接予測することを試みている。したがって、このような方法は、大量の画像群及びそれらの対応するピクセル深度が利用できるシーンに制限される。
【０００３】
他には、訓練中において、自動深度推定を画像再構築問題として扱うアプローチが進められている。人は、様々な手がかりのうち、遠近、既知の物体の既知のサイズに対するスケーリング、明暗の形態、オクルージョンなどの手がかりを生かし、単眼深度推定をうまく行う。この手がかりのトップダウン及びボトムアップの組み合わせは、全体シーンへの理解、及び深度を正確に推定する我らの能力を結び付けている。最近公表されたいくつかの研究により、訓練時にグラウンドトゥルース深度を必要としない新規のビュー合成及び深度推定のための、深層ネットワークベースの方法が提案されている。
【０００４】
J. Flynn, I. Neulander, J. Philbin及びN. Snavelyによる「DeepStereo: Learning To Predict New Views From The World's Imagery」（CVPR 2016）には、近くの隣接画像からピクセルを選択することにより新しいビューを生成する、DeepStereoと呼ばれる新規の画像合成ネットワークについて論議されている。訓練中において、彼らは画像のセットを選択し、彼らのそれぞれのカメラポーズを（動作からオドメトリ及び標準構造の組み合わせを用いて）計算し、その後、提供された近くの画像のアピアランスを予測するために畳み込みニューラルネットワーク（CNN）を訓練する。プレーンスイープ量に基づいて、隣接画像からカラーをサンプリングするために最も適した深度が選択される。テストのときに画像合成は小さい重ね合わせパッチで行われる。しかしながら、DeepStereoは、テストのときにいくつかの近くに位置する画像を必要とするので、単眼深度推定には適していない。
【０００５】
J. Xie, R. Girshick 及び A. Farhadiによる「Deep3d: Fully Automatic 2D-To-3D Video Conversion With Deep Convolutional Neural Networks」（ECCV 2016）

で論議されるDeep3D CNNも、訓練段階における新規のビュー合成問題を扱っており、彼らの目的は、両眼ステレオ画像ペアのコンテキストにおいて、左側入力画像（例えば、ソース画像）から対応する右側ビューを生成することである。コンピュータビジョンにおいてよく知られているように、両眼視差は、二つのステレオ画像内の同一特徴の座標の差、例えば、左右のカメラから見た物体の画像位置の差を指し、カメラ間の水平隔離（parallax）に起因するものである。Deep3Dは、立体視の二次元画像から深度情報を抽出するために両眼視差を用いる。画像再構築ロスを再度用いて、彼らの方法は、左側入力画像の各ピクセルに対して、可能性のある全ての視差にわたる分布を生成する。結果として生じる右側画像の合成されたピクセル値は、各視差の確率で重み付けされた、左側画像からの同じ走査線上のピクセルの組み合わせである。彼らの画像形成モデルの欠点は、候補視差値の数の増加がアルゴリズムのメモリ消費を非常に増加させて、彼らのアプローチを高出力解像度へスケーリングすることが難しいという点にある。

【０００６】

Deep 3Dと同様に、R. Garg, V. Kumar BG及びI. Reidによる「Unsupervised CNN For Single View Depth Estimation: Geometry To The Rescue」（ECCV 2016）には、訓練段階における両眼ステレオ画像ペアに基づいて、画像再構築ロスを用いる単眼深度推定のためにCNNを訓練することについて論議されている。しかし、Garg らによる画像形成モデルの記述は、完全微分可能なものではなく、訓練を最適なものにはできない。それを補うため、彼らは、ロスを線形化するテイラー近似を行い、最終結果の質を大きく増加させた。

【発明の概要】

【発明が解決しようとする課題】

【０００７】

望まれるのは、深度推定のための上記の深層CNNベースシステムの全ての制約に対処し、かつ最終結果の品質を著しく向上させる、改善されたネットワークアーキテクチャである。

【課題を解決するための手段】

【０００８】

本発明の態様は、添付の特許請求の範囲に述べられている。

【０００９】

一態様によれば、本発明は、カラー画像データから深度データを予測するための統計モデルを定義するデータを記憶し、入力された両眼ステレオペアの画像の少なくとも１つによってそのモデルを訓練し、その訓練は、入力された両眼ステレオペアの各画像について、その画像に適用されたときに他の画像の再構築を可能にする対応する視差値を予測すること、かつステレオペアの各画像に対して前記予測視差値間の整合性を高めるコスト関数に基づいてモデルを更新することにより行われる、コンピュータで実現される方法を提供する。

【００１０】

モデルの訓練は、ステレオペアの各画像に対して、対応する視差値に基づいて投影視差値を計算することをさらに含んでもよい。投影視差値は、ステレオペアの一方の画像に対して、第１画像の予測視差値をサンプリングし、かつサンプリングされたデータに他方の画像の予測視差値を適用することで計算されてもよい。コスト関数は、ステレオペアの各画像について計算された予測視差値と投影視差値との間の整合性を高めるための視差整合性構成要素を含んでもよい。

【００１１】

モデルの再構築モジュールは、ステレオペアの第１画像のサンプリングされた画像ピクセルをずらすために、対応する予測視差値を適用することで、ステレオペアの第２画像を再構築してもよい。コスト関数は、再構築画像と対応する入力画像との間の画像再構築誤差を最小にするための再構築アピアランスマッチング構成要素をさらに含んでもよい。サンプリングは、バイリニア補間を含んでもよい。

【００１２】
コスト関数は、対応する予測視差値における局所的な平滑化を促進するための平滑化構成要素をさらに含んでもよい。コスト関数は、視差整合性構成要素、平滑化構成要素、及び再構築アピアランスマッチング構成要素の重み付き和を実現してもよい。
【００１３】
統計モデルは、各処理ノードが少なくとも一つのパラメータ値を有する処理ノードの構造化配置を含む畳み込みニューラルネットワーク、すなわち CNN を含んでもよい。畳み込みニューラルネットワークは、コスト関数の逆伝播構成要素により訓練されてもよい。
【００１４】
モデルの訓練は、複数の空間解像度で入力画像データをアップサンプリング及びアップコンボリューションすること、及び各空間解像度で対応する視差値を予測することをさらに含んでもよく、モデルは、ステレオペアの各画像に対する各空間解像度で予測視差値間の整合性を高めるコスト関数に基づいて更新される。コスト関数は、空間解像度に応じて予測視差値間の整合性の重み付き強化を含んでもよい。
【００１５】
訓練画像の両眼ステレオペアは、既知のカメラ焦点長を有しかつ既知の基線距離だけ離れているそれぞれのカメラによって同時に撮像されてもよい。訓練画像の両眼ステレオペアは、修正されかつ時間的に整列されたステレオペアであってもよい。デジタル画像は、画像を撮像したそれぞれのカメラの属性を定義するメタデータで注釈付けされてもよい。
【００１６】
他の態様によれば、深度画像は、訓練されたモデルの視差予測モジュールを用いて、入力されたカラー画像から予測視差マップを生成すること、及び予測視差マップから対応する推定深度データを計算することにより、入力された単一カラー画像から生成されてもよい。カラー画像データは、カメラによって撮像されてもよい。モデルは、高解像度の画像を受信するように構成されてもよい。
【００１７】
有利なこととして、本発明は、深度データを必要としない代わりに深度を中間として合成するように訓練された全層畳み込みモデルを提供する。そのモデルは、既知のカメラ基線を有する、修正されたステレオ画像ペア間のピクセルレベルの対応関係を予測するように学習する。
【００１８】
さらに、実施形態は、
ネットワーク内に左右視差整合性制約を組み入れた新規の訓練ロスを用いてエンドツーエンド教師なし単眼深度推定を行うネットワークアーキテクチャ、
前述のアプローチの有効性を強調する、いくつかの異なる訓練ロス及び画像形成モデルの評価、及び
他の異なるデータセットによって一般化されるモデル
を提供する。
【００１９】
別の態様によれば、本発明は、シーン形状または存在するオブジェクトの種類の仮定がない単一入力画像のみの単眼深度推定のための教師なし深層ニューラルネットワークを提供する。特定の実施状況で利用できないかまたは得るのにコストがかかる可能性がある整列されたグラウンドトゥルース深度データの代わりに、本発明は、両眼ステレオデータを撮像できる簡易さを活用する。さらなる別の態様によれば、学習モジュールは、訓練中において各カメラビューから予測深度マップ間の整合性を高めるロス関数を実施して、予測を改善する。結果として生じる出力深度データは、訓練段階でグラウンドトゥルース深度情報を省略したにもかかわらず、完全に監督された基線より優れる。さらに、訓練されたモデルは、訓練中には見られなかったデータセットによって一般化されることができ、依然として視覚的に妥当な深度マップを生成することができる。
【００２０】

他の態様において、上述の方法を実行するように構成された装置及びシステムが提供される。さらなる態様において、プログラム可能なデバイスに上述の方法を実行させる機械読取可能な命令を含むコンピュータプログラムが提供される。

【図面の簡単な説明】

【００２１】

ここより、単なる例示として、以下に特定される図面を参照しながら本発明の実施形態について詳しく説明する。

【００２２】

【図１】本発明の一実施形態に係る画像処理システムの主要構成要素を示すブロック図である。

【図２】例示的なCNNの一部分を示す概略図である。

【図３Ａ】一実施形態に係る、単一画像深度予測CNNを訓練するための訓練モジュールにより実行される主要処理ステップを示すフロー図である。

【図３Ｂ】一実施形態に係る、単一画像深度予測CNNを訓練するための訓練モジュールにより実行される主要処理ステップを示すフロー図である。

【図４】一実施形態に係る、訓練の反復における例示的なCNNの処理及びデータ構成要素を概略的に示すブロックフロー図である。

【図５】一実施形態に係る、訓練されたCNNを用いて単一ソース画像から深度データを生成しかつ処理する例示的な処理を示すフロー図である。

【図６】一実施形態の機能のうち一つ以上を実施することができるコンピュータシステムの例を示す図である。

【発明を実施するための形態】

【００２３】

図1は、カラー画像データから深度データを予測しかつ処理するための例示的なシステム1を示すブロック図である。図に示すように、システム1は、画像処理システム3を含み、画像処理システム3は、カメラ7から撮像されたカラー画像データ（撮像されたビューでオブジェクトを形成するピクセルに対して RGB 値を表す RGB 画像など）を受信することができる深度データ生成器モジュール5を有する。デジタル画像は、画像を撮像したそれぞれのカメラの属性を定義するメタデータで注釈付けされることができる。深度データ生成器モジュール5は、受信された単一ソース画像のカラー画像データから予測両眼視差マップを直接生成するために、訓練された畳み込みニューラルネットワーク（CNN）モジュール11の視差予測器9を用いる。生成された両眼視差値は、ソース画像がキャリブレーションされた両眼ステレオカメラペアによって撮像されたステレオ画像ペアのうちの一つであるとみなした場合における、撮像されたソース画像内の検出されたオブジェクトまたは特徴の画像位置と、対応する概念的な両眼立体視におけるオブジェクトまたは特徴の予測画像位置との差を表す。深度データ生成器モジュール5は、視差予測器9により出力された両眼視差マップから深度情報を計算する。

【００２４】

CNN11は、処理ノードの動的構造化配置を含み、各ノードは対応する重みパラメータを有する。CNN11を定義する構造及び重みは、訓練段階において訓練モジュール13により更新される。この実施形態では、CNN11の処理ノードは、以下の3つの主要構成要素で構成される。

－以下のことを行うノード及び層を含むエンコーダ12：入力画像データを処理し、かつ入力画像内のオブジェクトまたは特徴を示す符号化されたデータを出力する。

－以下のことを行うノード及び層を含むデコーダ14：エンコーダ12からの符号化されたデータを処理し、アップコンボリューション及びアップサンプリングすることで、より大きな空間解像度のスケールされたデータを出力し、予測視差マップ（例えば、符号化されたデータの入力によって、視差予測器9により出力された視差マップ）を出力し、予測視差マップを入力画像データに適用することにより投影ビューを出力する。

－以下のことを行うノード及び層を持つロスモジュール19：CNN11を更新するために

用いられる訓練ロスを計算する。訓練ロスは、デコーダ14によって出力された視差マップから計算される視差平滑性項及び左右視差整合性コスト項、並びに投影ビューと対応する入力ビューとの比較から計算されたアピアランスマッチングコスト項を含む。

【００２５】

以下に詳しく説明されるように、訓練モジュール13は、訓練画像のデータベース17などから検索された両眼ステレオ画像ペア15に基づいて畳み込みニューラルネットワーク（CNN）モジュール11を訓練する。両眼ステレオ画像ペア15は、既知のカメラ焦点長及び既知の基線距離を有するそれぞれの両眼ステレオカメラにより同時に撮像された左側ビュー15a及び右側ビュー15bを含む。それによって、視差予測器9により出力された予測両眼視差値から深度データを計算することができる。訓練モジュール13は、CNN11モジュールのロスモジュール19によって実現されるロス関数を最適化し、結果として、単一ソース画像のカラーピクセル値から直接的に予測両眼視差マップを正確かつ効果的に生成するために視差予測器9を訓練する。

【００２６】

CNNモジュール11、訓練モジュール13及び深度データ生成器モジュール5は、単一モジュールに結合されるか、またはさらに複数のモジュールに分割されてもよく、画像処理モジュール3は、訓練されたCNNモジュール11のモデルデータを記憶するためのメモリ21などの追加的な構成要素を含んでもよいことを理解されたい。システム1は、計算システム／デバイスで一般的に見られる他の構成要素、副構成要素、モジュール及びデバイスといった、明確な説明のために図1に示していないものを含むことができる。

【００２７】

画像処理システム3により出力された深度情報は、さらなるデータ処理のために、一つ以上の深度データ処理モジュール23に提供されてもよい。深度データ処理モジュール23は、処理された深度データに基づいて、データ及び／または制御信号を出力デバイス25に出力するように構成されてもよい。深度データ処理モジュールの性質及び配置は、システム1の実施状況によって異なる。純粋に例示的な具体的実施形態としては、以下の通りである：コンピュータグラフィックスにおける合成オブジェクトの挿入に関連した、撮像された画像データからの深度マップの予測；コンピュータ写真における合成被写界深度の特定；ロボット把持のための制御命令の生成；人体ポーズ推定における手がかりとしての深度の出力；ヒューマンコンピュータインタラクションにおける手のポーズに対する強力な手がかりの特定；フィルムビデオデータの2Dから3Dへの自動変換；自律走行自動車の低コスト障害物回避センサ；手術のためのスモールフォームファクタ、シングルカメラ、深度感知、内視鏡；シングルカメラ3D再構築；VRヘッドセットのための改善されたポーズ推定；視覚障碍者のための障害物回避及び経路マッピング；物体計測のためのサイズ及び体積推定。訓練データ17は、特定の実施状況によって異なるビューのステレオ画像ペア15を含むことができることを理解されたい。

【００２８】

図2は、本実施形態による例示的なCN のデコーダ14及び訓練ロスモジュール19の部分を示す概略図である。CNN11の例示的な層は、以下の表1に示す通りであり、N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy及びT. Broxによる「A Large Dataset To Train Convolutional Networks For Disparity, OpticalFlow, And Scene Flow Estimation」（CVPR 2016）からの全層畳み込みアーキテクチャに基づいているが、グラウンドトゥルース深度データを必要とせずにネットワークを訓練することを可能にするいくつかの変更を含むように対応されている。図示の例示では、CNNは、訓練段階でシステムにより学習された3100万個のパラメータで構成されており、ここで、「k」はカーネルサイズであり、「s」はストライドであり、「チャンネル」は各層の入力及び出力チャンネルの数であり、「in」及び「out」はそれぞれの入力画像に対する各層の入力及び出力縮小率であり、「入力」は各層の入力に対応し、ここで「＋」は連続を意味し、「＊」は対応する層の２倍アップサンプリングに対応する。

【表１】

エンコーダ層	k	s	チャンネル	in	out	入力層
conv1	7	2	3/32	1	2	left
conv1b	7	1	32/32	2	2	conv1
conv2	5	2	32/64	2	4	conv1b
conv2b	5	1	64/64	4	4	conv2
conv3	3	2	64/128	4	8	conv2b
conv3b	3	1	128/128	8	8	conv3
conv4	3	2	128/256	8	16	conv3b
conv4b	3	1	256/256	16	16	conv4
conv5	3	2	256/512	16	32	conv4b
conv5b	3	1	512/512	32	32	conv5
conv6	3	2	512/512	32	64	conv5b
conv6b	3	1	512/512	64	64	conv6
conv7	3	2	512/512	64	128	conv6b
conv7b	3	1	512/512	128	128	conv7

デコーダ層	k	s	チャンネル	in	out	入力層
upconv7	3	2	512/512	128	64	conv7b
iconv7	3	1	1024/512	64	64	upconv7+conv6
upconv6	3	2	512/512	64	32	iconv7
iconv6	3	1	1024/512	32	32	upconv6+conv5
upconv5	3	2	512/256	32	16	iconv6
iconv5	3	1	512/256	16	16	upconv5+conv4
upconv4	3	2	256/128	16	8	iconv5
iconv4	3	1	128/128	8	8	upconv4+conv3
disp4	3	1	128/2	8	8	iconv4
upconv3	3	2	128/64	8	4	iconv4
iconv3	3	1	66/64	4	4	upconv3+conv2+disp4*
disp3	3	1	64/2	4	4	iconv3
upconv2	3	2	64/32	4	2	iconv3
iconv2	3	1	34/32	2	2	upconv2+conv1+disp3*
disp2	3	1	32/2	2	2	iconv2
upconv1	3	2	32/16	2	1	iconv2
iconv1	3	1	18/16	1	1	upconv1+disp2*
disp1	3	1	16/2	1	1	iconv1

【００２９】

上記のように、CNN11は、エンコーダ12（conv1層～conv7b層を含む）及びデコーダ14（upconv7層～disp1層を含む）を含む。当技術分野で知られているように、デコーダ14は、より高解像度を扱うために、エンコーダの活性化ブロックからスキップ接続を実装することができる。図2では、Cは畳み込み接続、UCはアップコンボリューション接続、Sはバイリニアサンプリング接続、USはアップサンプリング接続を指す。本例示的実施形態では、視差予測は異なる４つのスケール（disp4～disp1とラベル付け）で出力され、それらは後続の各スケールよりも空間解像度が増加している。ネットワークを訓練するとき、下付きsで表されるように、各出力スケールで各入力画像ビュー（例えば、左側及び右側ビュー）に対して2つの視差マップが予測される。一方の視差マップは、層への入力に基

づいてアラインされており（例えば、左から右への視差マップd^rは、左側ビューの符号化されたデータに基づいてアラインされている）、他方の視差マップは、その対応するステレオパートナについてアラインされる（例えば、右から左への投影視差マップd^l（d^r）は、対応する投影右側ビューに基づいてアラインされる）。デコーダ14及びロスモジュール19による処理は、異なる4つの出力スケールのそれぞれで繰り返される。

【００３０】

重要な利点は、訓練されたシステム3が、両眼カメラの両方から視差を予測し、かつそれらを互いにより整合するようにすることで、優れた深度マップを生成することである。左側ビューからのピクセルを用いて右側ビューを生成することで、右側ビューとアラインされた視差マップが得られる。（逆もまた同様）。訓練モジュール13は、予測視差マップのソース入力画像（本実施形態では左側ビュー15a）へのアラインメントを最適化することを目的とする。訓練中において、訓練モジュール13は、左右のステレオ画像15a、15bの両方にアクセスして、左から右への及び右から左への視差マップを推定し、かつそれぞれの推定視差マップから左から右への及び右から左への対応する投影視差マップを特定し、かつ視差マップ間の整合性を高めるようにCNN11を訓練する。訓練モジュール13のさらなる最適化の目標は、画像再構築誤差を最小にするためにピクセルをずらすことができる視差マップを学習することによって、対応する左側及び右側ビューを再構築するように、CNN11を訓練することである。このようにして、校正された両眼カメラペアから訓練画像が与えられると、画像処理システム3は、他のビューが与えられた画像を再構築できる関数を学習し、そのようにすることで、画像となるシーンの形状の予測または推定を可能にする訓練されたモデル（例えば、CNN11）を生成する。単一訓練画像I（例えば、訓練ステレオ画像ペア15の左側ビュー15a）が与えられると、画像処理システム3は、ピクセル毎のシーン深度を予測できる関数である $d^=f(I)$ を学習し、深度推定を訓練中における画像再構築問題として扱う。

【００３１】

以上、実施形態の画像処理システム3の構成要素を形成する部分について概要説明を行った。対応するグラウンドトゥルース深度情報形式などの教師を必要とせずにステレオ画像ペアでのみCNN11を訓練できるようにする、一実施形態による、単一画像深度予測CNN11を訓練する処理について、図3のフロー図を参照しながら、これら構成要素の動作のより詳しい説明を行う。このフローチャートの様々なステップを順に示し説明するが、ステップの一部または全部は、異なる順で実行されてもよく、結合または省略されてもよく、ステップの一部または全部は並列して実行されてもよいことを理解されたい。さらに、一つ以上の例示的な実施形態において、以下の一つ以上のステップは省略されるか、繰り返されるか、及び／または異なる順で実行されてもよい。

【００３２】

本発明の実施形態による、CNN11の処理ノード及び層の例示的な構造化配置を概略的に示すブロックフロー図である図4も参照することができる。図2に示すアップコンボリューション（UC）及びアップサンプリング（US）層は、簡潔にするため図4からは省略されているが、UC及びUS層からのスケーリング済み出力は、各予測視差及びそれぞれの計算されたコスト要素に下付きsで表されることを理解されたい。

【００３３】

図3に示すように、単一ペアの訓練画像15に対する訓練処理の反復は、CNN11が入力ステレオペアのうち一方のビュー（この実施形態では左側ビュー）のカラー画像データを受信するステップS3-1(L）で始まる。この実施形態では、CNN11は、ステップS3-1（R）で右側ビューのカラー画像データも受信する。訓練モジュール13は、メモリ17に記憶された訓練データから、校正されたステレオペアの対応する左右のカラー画像でありかつ同時に撮像された二つの画像I^l及びI^rを検索することができ、CNN11の一つ以上の入力ノード（図示せず）に画像データを渡すことができる。CNN11は、複数の訓練画像ペアを、好適に同時に受け取って処理するように構成されてもよいことを理解されたい。必ずしも必要ではないが、好ましくは、当技術分野で知られているように、ステレオ画像ペア15は修

正され、それによって、定義された変換処理で画像が共通画像平面に投影される。

【００３４】

ステップS3-3で、左側ビューの入力画像データは、例えば、入力画像内の識別されたオブジェクトまたは特徴の複素特徴ベクトルを表す符号化された入力データを生成するために、エンコーダ12の畳み込み層を通過する。左側ビュー画像15aから深度を直接予測しようとする代わりに、CNN11は、対応するフィールドを探索するように訓練される。この実施形態において該フィールドは、左側ビュー画像15aに適用されたときに、CNN11の右側ビュープロジェクタ415aが投影右側ビュー画像を再構築できるようにする左から右への予測視差マップ（d^r）である（逆もまた同様）。したがって、ステップS3-5で、ステップS3-3において出力された符号化されたデータは、現在の構造及び重みに基づいて左から右への予測視差マップ（$d^r{}_s$）のデータ値を出力する左側ビュー視差予測器307aの処理ノードを通過する。後述するように、CNN11は、入力された両眼ステレオペアの各画像に対して対応する視差値を予測し、かつステレオペアの各画像に対して予測視差値間の整合性を高くするコスト関数に基づいてCNN11を更新することによって、入力データから視差マップを予測するように訓練される。したがってステップS3-5で、符号化されたデータはまた、現在の構造及び重みに基づいて右から左への予測視差マップ（$d^l{}_s$）のデータ値を出力する右側ビュー視差予測器307bの処理ノードを通過する。

【００３５】

任意で、ステップS3-7で、ロスモジュール13のL-〉R視差平滑化ノード413aにより、左から右への予測視差マップ（d^r）から、左から右への視差平滑化コスト$(C^r{}_{ds})_s$が計算されてもよい。同様に、ステップS3-7で、ロスモジュール13のR-〉L視差平滑化ノード413bにより、右から左への予測視差マップ（d^l）から右から左への視差平滑化コスト$(C^l{}_{ds})_s$が計算されてもよい。訓練ロス関数の計算された平滑化コスト要素は、ぞれぞれの予測視差マップを、視差勾配∂dについてのL1ペナルティで局所的に平滑化する。例えば、左の予測視差マップd^lから計算された平滑化コストは、以下のように式化される。

【数１】

$$C_{ds}^{l} = \frac{1}{N}\sum_{i,j}\left|\partial_x d_{ij}^{l}\right| e^{-\eta\left\|\partial_x I_{ij}^{l}\right\|} + \left|\partial_y d_{ij}^{l}\right| e^{-\eta\left\|\partial_y I_{ij}^{l}\right\|} \tag{1}$$

ここで、ηは1.0と設定することができる。画像勾配で奥行不連続性がたびたび発生するので、この平滑化コストは、対応する画像勾配∂Iを用いてエッジを意識した項で重み付けされうる。

【００３６】

ステップS3-9(L)で、R-〉L視差プロジェクタ409aは、左から右への予測視差マップ（$d^r{}_s$）のデータ値をサンプリングし、右から左への投影視差マップ$(d^l(d^r)_s)$を生成するために、右から左への予測視差マップ$(d^l{}_s)$をサンプリングされたデータに適用する。明確にするために、左から右への予測視差値の処理は、(L)と示すステップを参照しながら説明される。対応する番号の処理ステップは、(R)と示すように、右から左への視差値に同様に反映されたものであることを理解されたい。この実施形態では、視差プロジェクタ409は、M. Jaderberg, K. Simonyan, A. Zisserman及びK. Kavukcuogluによる「Spatial Transformer Networks」（NIPS 2015）などから当分野で知られているように、STN（spatial transformer network）からの画像サンプラに基づいて、視差マップを用いて入力データをサンプリングするために、画像サンプリング機能を実装する。STNは、出力ピクセルが4入力ピクセルの加重和であるバイリニアサンプリングを用いる。前述のXieら及びGargらによるアプローチとは対照的に、この実施形態で用いられるバイリニアサンプラは、局所的に完全微分可能であり、CNN11の全層畳み込みアーキテクチャにシームレスに統合される。これは、CNN11が最適化コスト関数の如何なる単純化または近似値も必要としないことを意味する。

【００３７】
より確実な結果を生むために、CNN11は、ネットワークの畳み込みロスモジュール13部分への入力として左側ビュー画像データ15aのみに基づいて、左右の画像視差の両方を予測するように訓練される。したがって、ステップS3-9(L)で、CNN11の投影された右視差予測器ノード409aは、ステップS3-5(L)で左側ビュー視差予測器ノード407aにより出力された左の予測視差マップ（d^l）に基づいて、投影された右視差マップ（$d^l(d^r)$）を出力する。一貫性を確実にするため、ロスモジュール13は、モデル11の一部としてL1左右視差整合性ペナルティを含む。このコストは、予測された左側ビュー視差マップ（d^l）を、投影された右側ビュー視差マップ（$d^r(d^l)$）と等しくするために提供される。したがって、ステップS3-11(L)で、L-R視差整合性ロスノード411aは、以下のように左整合性コストを計算する。
【数２】

$$C_{lr}^{l} = \frac{1}{N}\sum_{i,j}\left|d_{ij}^{l} - d_{ij+d_{ij}^{l}}^{r}\right| \qquad (2)$$

【００３８】
ステップS3-13（L）において、CNN11の粗密スケーラ405aは、スケールs_1～s_nで左側ビューのスケーリング済み画像データを生成しかつ出力する。本例示的実施形態ではn=4である。各スケールsに対して、左側ビューの対応するスケーリング済み画像データ（I^l_s）は、処理のためにデコーダ14の右側ビュープロジェクタ415aに渡される。ステップS3-15(L)で、右側ビュープロジェクタ415aは、スケーリング済み左側ビュー画像（I^l_s）からピクセルをサンプリングすることによって投影隣接ステレオ画像を生成する。この実施形態では、ビュープロジェクタ415は、入力視差マップを用いて入力データをサンプリングするために、上述のようなSTN（spatial transformer network）からの画像サンプラをさらに実装する。
【００３９】
任意で、ステップS3-15(L)で、CNN11の右側ビュープロジェクタ415aは、左から右への予測視差（d^r）を入力されたスケーリング済み左側ビュー画像データ（I^l_s）に適用することにより、左側ビュー投影画像を再構築することができる。この処理は、以下のように式化することができる。
【数３】

$$arg\ min_{d_r} \left\| I^r - I^l(d^r) \right\| \qquad (3)$$

ここで、dは、モデル11が予測するように訓練されたピクセルごとのスケーラ値である画像視差に対応する。再構築画像$I^l(d^r)$は、簡潔にするために、$\tilde{I}^r$と呼ぶ。左側ビュー投影画像は、ステップS3-13(R)及びS3-15(R)において、右から左への予測視差マップ（d^l）を入力されたスケーリング済み右側ビュー画像データ（I^r_s）に適用することにより、同様に生成されることができる。
【００４０】
ステップS3-17(L)で、アピアランスマッチングコストは、Rアピアランスマッチングロスノード417aにより、L1項及びSSIM(single scale Structured Similarity)項の組み合わせとして計算することができ、側光の入力画像I^l_{ij}とその再構築$\tilde{I}^l_j$との間の画像再構築コストは、
【数４】

$$C_{ap}^{l} = \frac{1}{N}\sum_{i,j} \propto \frac{1-\mathrm{SSIM}(I_{ij}^{l}, \tilde{I}_{ij}^{l})}{2} + (1-\propto)\left\|I_{ij}^{l} - \tilde{I}_{ij}^{l}\right\| \quad (4)$$

ここで、Nは画像のピクセル数である。例示的な実施形態では、ガウシアンフィルタの代わりに3x3ブロックフィルタを有する単純化SSIMが用いられ、αは0.85と設定される。SSIM項の計算法は、Z. Wang, A. C. Bovik, H. R. Sheikh及びE. P. Simoncelliによる「Image Quality Assessment: From Error Visibility To Structural Similarity」（Transactions on Image Processing 2004）などから分かるように、それ自体が知られているので、さらなる説明を必要としない。左アピアランスマッチングコスト（C^{l}_{ap}）は、ステップS3-17(R)において、左側ビュープロジェクタ415bにより出力された投影左側ビュー、及び、スケーラ405により出力された対応するスケーリング済み左側ビュー画像から、同様に計算されることができる。

【００４１】

左側ビューカラー画像15a及び右側ビューカラー画像15bがCNN11を通過すると、ステップS3-19において、ロスモジュール13の訓練ロスノード419は、現在のスケールで訓練画像のステレオペアに対する訓練ロスを計算する。本実施形態において、スケーリング済み訓練ロスは、ステップS3-7(L）及び(R)で出力された視差平滑化コスト、ステップS3-11(L)及び(R)で出力された視差整合性コスト、及びステップS3-17(L)及び(R)で出力されたアピアランスマッチングコストの重み付き組み合わせとして計算される。この計算された3つのコスト項の重み付き組み合わせは、以下のように式化することができる。

【数５】

$$C_{s=}\alpha_{ap}\left(C_{ap}^{l} + C_{ap}^{r}\right) + \alpha_{ds}\left(C_{ds}^{l} + C_{ds}^{r}\right) + \alpha_{lr}\left(C_{lr}^{l} + C_{lr}^{r}\right) \quad (5)$$

ここで、C_{ap}は、再構築画像が対応する訓練入力に類似するようにし、C_{ds}は、視差を平滑化させ、C_{lr}は、左右の画像からの予測視差が整合するように試みる。各主要項は、左右両方の画像の変形を含む。3つの訓練コスト要素を全て含むこの例示的な実施形態では、左側ビュー画像15aは、常にCNN11を通過する。訓練中において、訓練モジュール13が対応する右側ビュー画像15bへのアクセスを有するので、CNN11はその参照フレームにおける視差マップを予測することもできる。アピアランスマッチングコスト要素が実装されていない場合、右側ビュー画像データはCNN11を通過する必要がないことを理解されたい。

【００４２】

ステップS3-21において、CNN11のデコーダ14は、前述のステップS3-3で論議されたように、次のスケールのためのスケール済み訓練ロスを計算するための後続の構造化された処理ノードのセットへの入力として、現在のスケールでエンコーダ12により出力されたデータのアップコンボリューション、及び視差予測器407により出力された予測視差マップのアップサンプリングを実行する。スケーリング済み訓練ロスが各所定スケールに対して計算された後、ステップ3-23において最終的な総ロスが、ロスモジュール13の合計ノード421により個々のスケーリング済みロスC_sの重み付き和として、以下のように計算される。

【数６】

$$C = \sum_{s=1}^{4} \lambda_s C_s \quad (6)$$

ここで、λ_sは、訓練モジュール13が訓練中において異なる出力スケールの相対的な重要度で重み付けされるするようにする。

【００４３】
例示的な実施形態では、異なるロス構成要素の重み付けは、α_{ap}=1及びα_{lr}=1と設定される。可能性のある出力視差は、スケール済みのシグモイド非線形性を用いて、0～d_{max}の間に制限される。ここで、d_{max}は所与の出力スケールにおける画像幅の0.3倍である。マルチスケール出力の結果として、隣接ピクセルの一般的な視差は、（CNN11が出力を2倍アップサンプリングしているため）各スケール間で2倍異なる。これを修正するために、訓練モジュール13は、各レベルで同等の平滑化を得るために、各スケールに対して視差平滑化項α_{ds}をrでスケーリングすることができる。したがって、α_{ds}=0.1/rである。ここで、rはCNN11に渡される入力画像の解像度に対する、対応する層の縮小率である（表1より）。

【００４４】
ステップS3-25で、訓練モジュール13は、ステップS3-21において合計ノード421により計算された最終的な総訓練ロスの重み付き構成要素を逆伝播することにより、CNN11を更新する。CNN11における非線形性に対して、一般的に用いられる正規化線形ユニット（ReLU）の代わりに指数線形ユニットを用いることもでき、両方とも当技術分野で知られている。例示的な実施形態において、CNN11は、D. Kingma及びJ. Baによる「Adam: A method for stochastic optimization」（arXiv preprint, arXiv: 1412.6980， 2014）に記載の技術に基づいて、最初から50エポック訓練される。ここで、β_1=0:9、β_2=0:999、及び$\varepsilon=10^{-8}$である。初期学習速度は$\lambda=10^{-4}$であり、最初の30エポックでは一定に保たれ、その後は終了するまで10エポックごとに半分になる。訓練モジュール13は、低解像度画像スケールから最初に最適化される段階的更新スケジュールを用いてCNN11を更新するように構成されてもよいことを理解されたい。しかし、発明者らは、4つのスケールを全て一度に最適化することが、さらに有利により安定した収束をもたらすことに想到した。同様に、異なる重み付けが不安定な収束をもたらす場合は、各スケールロスの同一の重み付けを用いることができる。

【００４５】
図6は、一実施形態に係る、訓練されたCNN11を用いて単一ソース画像から深度データを生成しかつ処理する例示的な処理を示すフロー図である。ステップS6-1で、単一ソース画像のカラー画像データが、例えばカメラ7から深度データ生成器5に受信される。ステップS6-3で、訓練されたCNN11を通過するシングルフォワードを用い、左画像に対する最も微細なスケールレベルでの視差d^rが、訓練されたL-Rビュー視差予測器407aにより、予測視差マップとして出力される（表1のdisp1に対応する）。CNN11内の粗密スケーラ405によるアップサンプリングの結果として、出力予測視差マップは、入力画像と同じ解像度である。右から左への視差d^lは、深度データ生成段階では用いられないことを理解されたい。

【００４６】
ステップS6-5において、深度データ生成器5は、ステップS6-3において出力された予測視差マップから計算された、ソース画像の各ピクセルに対する予測深度データからなる深度画像を作成する。訓練データ15を撮像するためのステレオカメラ間の基線距離b、及び関連するカメラ焦点長fが与えられると、深度データ生成器5は、以下のように、予測視差から推定深度値を復元することができる。

【数７】

$$\hat{d} = b\frac{f}{d} \tag{7}$$

【００４７】
ステップS6-7で、深度画像は、システム1の具体的な実施状況に応じて処理されるように、深度データ処理モジュール23に渡される。

コンピュータシステム

【００４８】

画像処理システム3、及び／または画像処理システム3の個々のモジュールなど、本明細書に記載されるエンティティは、図6に示すようなコンピュータシステム1000などのコンピュータシステムにより実施されることができる。本発明の実施形態は、そのようなコンピュータ1000による実行のためのプログラム可能なコードとして実施されることができる。この説明を読んだ後、当業者であれば、他のコンピュータシステム及び／またはコンピュータアーキテクチャを用いての本発明の実施方法が明確になるであろう。

【００４９】

コンピュータシステム1000は、パーソナルコンピュータ、ラップトップ、コンピューティング端末、スマートフォン、タブレットコンピュータなどであってもよく、プロセッサ1004などの一つ以上のプロセッサを含む。プロセッサ1004は、任意のタイプのプロセッサであってもよく、特殊な目的または汎用のデジタルシグナルプロセッサを含むが、これに限定されない。プロセッサ1004は、通信基盤1006（例えば、バスまたはネットワーク）に接続される。この例示的なコンピュータシステムに関して、様々なソフトウェア実施が説明される。この説明を読んだ後、当業者であれば、例えば集積された入力及び表示構成要素を有する携帯電子装置を用いるなど、他のコンピュータシステム及び／またはコンピュータアーキテクチャを用いての本発明の実施方法が明確になるであろう。

【００５０】

コンピュータシステム1000は、一つ以上の入力装置1005に接続されたユーザ入力インタフェース1003、及び一つ以上の表示装置1009に接続された表示インタフェース1007も含む。入力装置1005は、例えば、マウスまたはタッチパッドのようなポインティング装置、キーボード、抵抗膜方式または容量方式タッチスクリーンのようなタッチスクリーンなどを含むことができる。この説明を読んだ後、当業者であれば、他のコンピュータシステム及び／またはコンピュータアーキテクチャを用いての本発明の実施方法が明確になるであろう。

【００５１】

コンピュータシステム1000は、主記憶装置1008、好ましくはRAM（random access memory）も含み、二次記憶装置も含むことができる。二次記憶装置1010は、例えば、ハードディスクドライブ1012、及び／または、フロッピーディスクドライブ、磁気テープドライブ、光ディスクドライブで表されるリムーバブルストレージドライブ1014などを含むことができる。リムーバブルストレージドライブ1014は、周知の方法により、リムーバブル記憶部1018から読み取るか、またはリムーバブル記憶部1018に書き込む。リムーバブル記憶部1018は、リムーバブルストレージドライブ1014に読み取られ、かつ書き込まれる、フロッピーディスク、磁気テープ、光ディスクなどで表される。リムーバブル記憶部1018は、コンピュータソフトウェア及び／またはデータを記憶しているコンピュータ使用可能な記憶媒体を含むことが理解されるであろう。

【００５２】

代替的な実施形態では、二次記憶装置1010は、コンピュータプログラムまたは他の命令をコンピュータシステム1000にロードさせる他の同様の手段を含むことができる。そのような手段は、例えば、リムーバブル記憶部1022及びインタフェース1020を含むことができる。そのような手段の例示は、プログラムカートリッジ及びカートリッジインタフェース（以前はビデオゲーム機で見られたものなど）、リムーバブルメモリチップ（EPROM、またはPROM、またはフラッシュメモリなど）、関連するソケット、及び、ソフトウェア及びデータをリムーバブル記憶部1022からコンピュータシステム1000に転送させる他のリムーバブル記憶部1022及びインタフェース1020を含むことができる。または、コンピュータシステム1000のプロセッサ1004を用いて、プログラムを実行し、及び／またはリムーバブル記憶部1022からデータをアクセスすることができる。

【００５３】

コンピュータシステム1000は、通信インタフェース1024も含むことができる。通信イ

ンタフェース1024は、コンピュータシステム1000と外部装置との間でソフトウェア及びデータを転送させる。通信インタフェース1024の例示は、モデム、ネットワークインターフェース（イーサネット（登録商標）カードなど）、通信ポート、PCMCIA（Personal Computer Memory Card International Association）スロット及びカードなどを含む。通信インタフェース1024を介して転送されるソフトウェア及びデータは、信号1028の形態であり、通信インタフェース1024により受信されることが可能である電子、電磁、光、または他の信号であり得る。これらの信号1028は、通信経路1026を介して、通信インタフェース1024に提供される。通信経路1026は、信号1028を運び、ワイヤまたはケーブル、光ファイバ、電話線、ワイヤレスリンク、携帯電話リンク、無線周波数リンク、または任意の他の適切な通信チャンネルを用いて実施されることができる。例えば、通信経路1026は、チャンネルの組み合わせを用いて実施されることができる。

【００５４】

「コンピュータプログラム媒体」及び「コンピュータ使用可能媒体」という用語は、一般的に、リムーバブルストレージドライブ1014、ハードディスクドライブ1012にインストールされたハードディスク、及び信号1028などの媒体を示すことに用いられる。これらのコンピュータプログラム製品は、コンピュータシステム1000にソフトウェアを提供するための手段である。しかし、これらの用語は、本明細書に開示されたコンピュータプログラムを具現する信号（電気信号、光信号、電磁気信号など）も含むことができる。

【００５５】

コンピュータプログラム（コンピュータ制御ロジックとも呼ぶ）は、主記憶装置1008及び／または二次記憶装置1010に記憶される。また、コンピュータプログラムは、通信インタフェース1024を介して受信されることができる。このようなコンピュータプログラムが実行されると、コンピュータシステム1000が、本明細書に説明されるように本発明の実施形態を実施することができるようにする。したがって、このようなコンピュータプログラムは、コンピュータシステム1000の制御装置を表す。この実施形態がソフトウェアを用いて実施された場合、いくつかの例示を提供するためにソフトウェアは、コンピュータプログラム製品1030に記憶され、リムーバブルストレージドライブ1014、ハードディスクドライブ1012、または通信インタフェース1024を用いてコンピュータシステム1000にロードされることができる。

【００５６】

代替的実施形態は、ハードウェア、ファームウェア、ソフトウェア、またはそれらの任意の組み合わせにおける制御ロジックとして実施されることができる。例えば、訓練されたCNNモジュール11は、画像処理システムにおける構成要素としてインストールするための単独エンティティとしてハードウェア及び／またはソフトウェアで実施されることができ、さらに、訓練モジュール及び／または深度データ生成器機能を含むことができる。

【００５７】

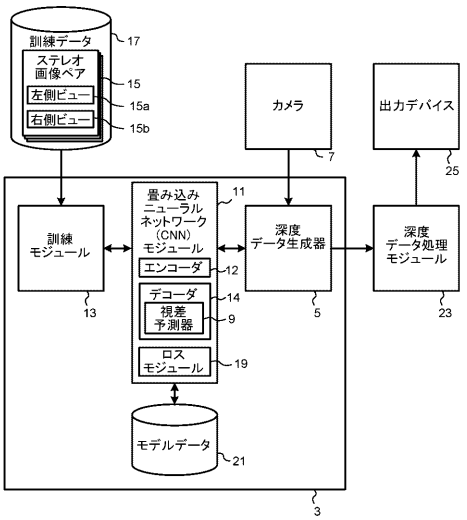
本発明の実施形態は、単に例示として本明細書に記載されたものであり、かつ本発明の範囲から逸脱することなく、様々な変更及び修正が可能であることが理解されるであろう。例えば、上述の実施形態は、訓練された統計モデルを深層畳み込みニューラルネットワークとして実施する。当業者に理解されるように、訓練処理の根本的な態様は、ランダムフォレスト及びその派生など、予測された深度マップを生成するために、画像データの処理に適した他の形態の統計モデルに適用可能であり得る。

【００５８】

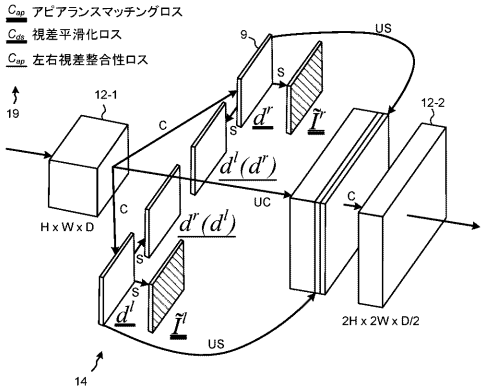
本明細書の「一実施形態」という言及は、必ずしも全て同じ実施形態を示すものではなく、他の実施形態の相互排他的な別のまたは代替の実施形態も含む。特に、上述の実施形態の態様を結合して、さらなる実施形態を形成することもできることを理解されたい。同様に、他の実施形態ではなく、いくつかの実施形態により示されることができる様々な特徴が説明される。それでもなお特許請求の範囲の範疇に含まれる、よりさらなる別の実施形態が考えられ得る。

【図面】

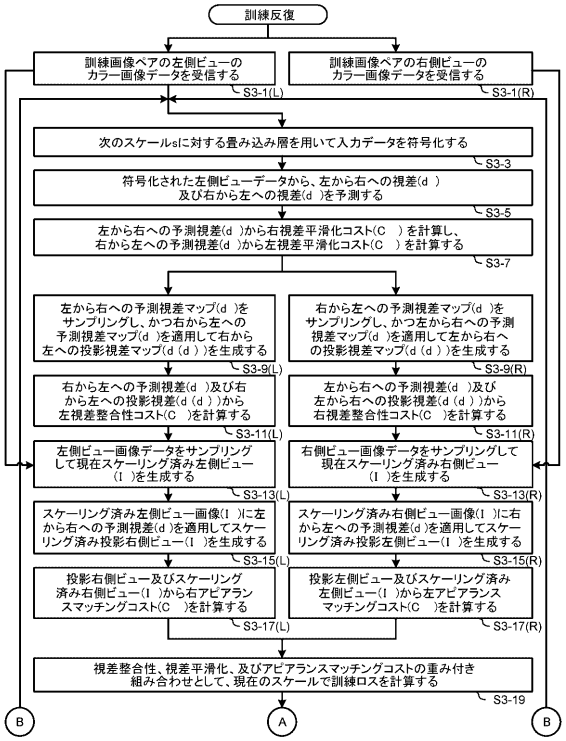
【図１】



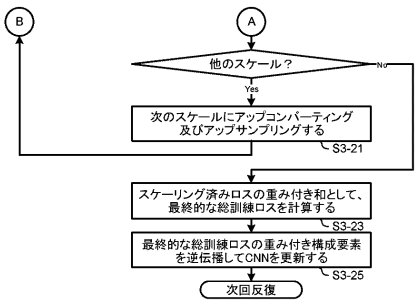
【図２】



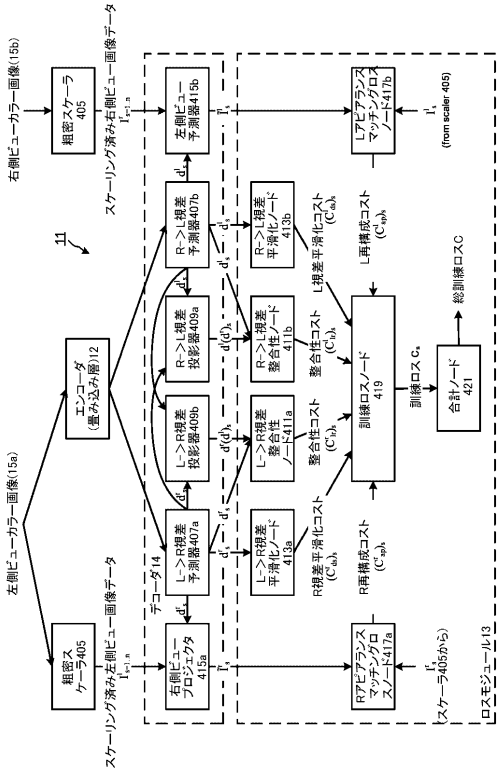
【図３Ａ】



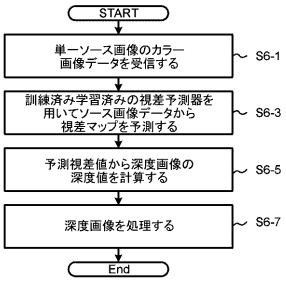
【図３Ｂ】



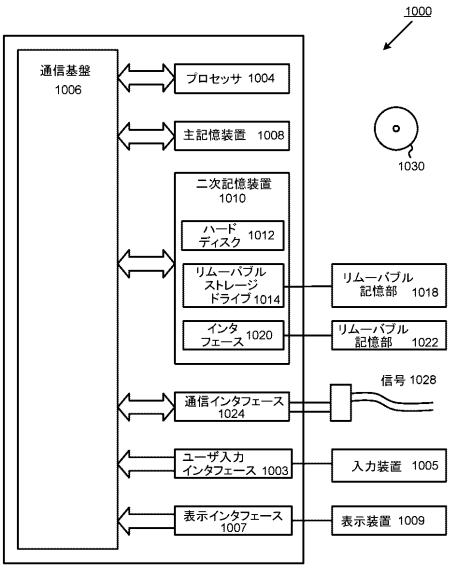
【図４】


右側ビューカラー画像(15b)
粗密スケーラ 405
スケーリング済み右側ビュー画像データ
左側ビューカラー画像(15a)
エンコーダ（畳み込み層）12
デコーダ14
粗密スケーラ405
スケーリング済み左側ビュー画像データ
左側ビュー予測器415b
R->L視差予測器407b
R->L視差投影器409a
L->R視差投影器409b
L->R視差予測器407a
右側ビュープロジェクタ415a
R->L視差平滑化ノード413b
R->L視差整合性ノード411b
L->R視差整合性ノード411a
L->R視差平滑化ノード413a
Lアピアランスマッチングロスノード417b
Rアピアランスマッチングロスノード417a
L再構成コスト
L視差平滑化コスト
整合性コスト
R視差平滑化コスト
R再構成コスト
訓練ロスノード419
合計ノード421
総訓練ロス
(from scaler 405)
(スケーラ405から)
ロスモジュール13

【図５】


START
単一ソース画像のカラー画像データを受信する
S6-1
訓練済み学習済みの視差予測器を用いてソース画像データから視差マップを予測する
S6-3
予測視差値から深度画像の深度値を計算する
S6-5
深度画像を処理する
S6-7
End

【図６】


1000
通信基盤 1006
プロセッサ 1004
主記憶装置 1008
二次記憶装置 1010
ハードディスク 1012
リムーバブルストレージドライブ 1014
インタフェース 1020
通信インタフェース 1024
ユーザ入力インタフェース 1003
表示インタフェース 1007
1030
リムーバブル記憶部 1018
リムーバブル記憶部 1022
信号 1028
入力装置 1005
表示装置 1009

フロントページの続き

(72)発明者　ゴダール クレメント
英国 ロンドン エスダブリュー８ ２エイチエックス ダヴィッドソンガーデンズ アルドウィンハウス フラット １１
(72)発明者　マック エイダ オシン
アメリカ合衆国 カリフォルニア州 ９１１０６ パサデナ アルペンストリート ２７６ ユニット ３
(72)発明者　ブロストウ ガブリエル
英国 ロンドン エヌ１２ ９ディーエス グローブロード ６５

審査官　岡本 俊威

(56)参考文献　特開２００５－１６５６１４（ＪＰ，Ａ）
Ravi Garg, Vijay Kumar B G, Gustavo Carneiro, Ian Reid，Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue，arxiv.org，米国，CORNELL UNIVERSITY，2016年03月16日，https://arxiv.org/pdf/1603.04992v1

(58)調査した分野　(Int.Cl.，ＤＢ名)
Ｇ０６Ｔ　７／５０－ ７／５９３