

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7640569号
(P7640569)

(45)発行日 令和7年3月5日(2025.3.5)

(24)登録日 令和7年2月25日(2025.2.25)

(51)国際特許分類

F I

G 0 6 T 7/55 (2017.01)

G 0 6 T 7/00 (2017.01)

G 0 1 B 11/245 (2006.01)

G 0 6 T 7/55

G 0 6 T 7/00

G 0 1 B 11/245

3 5 0 C

H

請求項の数 14 (全33頁)

(21)出願番号	特願2022-552548(P2022-552548)	(73)特許権者	514108838
(86)(22)出願日	令和3年3月5日(2021.3.5)		マジック リープ, インコーポレイテッド
(65)公表番号	特表2023-515669(P2023-515669 A)		M a g i c L e a p, I n c .
(43)公表日	令和5年4月13日(2023.4.13)		アメリカ合衆国 フロリダ 3 3 3 2 2 ,
(86)国際出願番号	PCT/US2021/021239		プランテーション, ウェスト サンライズ
(87)国際公開番号	WO2021/178919		ブルバード 7 5 0 0
(87)国際公開日	令和3年9月10日(2021.9.10)		7 5 0 0 W S U N R I S E B L V D
審査請求日	令和6年2月28日(2024.2.28)		, P L A N T A T I O N, F L 3 3 3
(31)優先権主張番号	62/985,773	(74)代理人	2 2 U S A
(32)優先日	令和2年3月5日(2020.3.5)		100078282
(33)優先権主張国・地域又は機関	米国(US)	(74)代理人	弁理士 山本 秀策
早期審査対象出願		(74)代理人	100113413
		(74)代理人	弁理士 森下 夏樹
		(74)代理人	100181674
			弁理士 飯田 貴敏

最終頁に続く

(54)【発明の名称】 マルチビューステレオのための疎点の三角測量および緻密化を学習することによる深度推定のためのシステムおよび方法

(57)【特許請求の範囲】

【請求項1】

マルチビュー画像からの場面内の特徴の深度を推定するための方法であって、前記方法は、

前記場面のアンカ画像と、前記場面の基準画像のセットとを含むマルチビュー画像を取得することと、

(1) 前記アンカ画像および各基準画像に関する記述子の個別の記述子フィールドを出力し、(i i) 相対的姿勢と併せて、前記アンカ画像内の疎着目点を検出し、代替視点から前記基準画像内の検索空間を決定し、(i i i) 中間特徴マップを出力する、共有RGBエンコーダおよび記述子デコーダを通して、前記アンカ画像および基準画像を通過させることであって、前記記述子デコーダは、記述子を訓練するように構成される学習アーキテクチャを利用する、ことと、

各基準画像の前記検索空間内の前記個別の記述子をサンプリングし、前記検索空間内の記述子を決定し、前記決定された記述子と前記アンカ画像内の前記疎着目点に関する記述子をマッチングさせることであって、そのようなマッチングされる記述子は、マッチングされた特徴点と称される、ことと、

特異値分解(SVD)を使用して、前記マッチングされた特徴点を三角測量し、3D点を出力することと、

前記3D点を疎深度エンコーダを通して通過させ、疎深度画像を前記3D点および出力特徴マップから作成することと、

深度デコーダが、前記疎深度エンコーダに関する前記出力特徴マップと、前記共有RGBエンコーダからの前記中間特徴マップとに基づいて、稠密深度画像を生成することを含む、方法。

【請求項2】

前記共有RGBエンコーダおよび記述子デコーダは、前記アンカ画像およびトランザクション画像の完全分解能に作用するように構成される完全畳み込みニューラルネットワークである、請求項1に記載の方法。

【請求項3】

前記共有RGBエンコーダからの前記中間特徴マップを第1のタスク特有のデコーダヘッドの中にフィードし、前記アンカ画像内の前記疎着目点の検出のための加重を決定し、着目点記述を出力すること

をさらに含む、請求項1に記載の方法。

【請求項4】

前記記述子デコーダは、U-Net様アーキテクチャを備え、前記決定された記述子と前記疎着目点に関する記述子をマッチングさせるために、微細および粗大レベル画像情報を融合させる、請求項1に記載の方法。

【請求項5】

前記検索空間は、前記基準画像内の個別のエピポーラ線および前記エピポーラ線の両側上の固定されたオフセットに制約され、前記エピポーラ線に沿った実行可能深度感知範囲内にある、請求項1に記載の方法。

【請求項6】

双線形サンプリングが、前記共有RGBエンコーダおよび記述子デコーダによって使用され、前記個別の記述子を前記記述子フィールド内の所望の点において出力する、請求項1に記載の方法。

【請求項7】

前記マッチングされた特徴点を三角測量するステップは、

空間軸を横断して、softmaxを算出し、相互相関マップを出力することによって、前記疎着目点の個別の2次元(2D)位置を推定することと、

softmaxargmax演算を実施し、節点の2D位置を対応する相互相関マップの質量中心として計算することと、

線形代数三角測量を前記推定された個別の2D位置から実施することと、

特異値分解(SVD)を使用して、3D点を出力することと

を含む、請求項1に記載の方法。

【請求項8】

クロスリアリティシステムであって、

ディスプレイシステムを有する頭部搭載型ディスプレイデバイスと、

前記頭部搭載型ディスプレイデバイスと動作可能に通信するコンピューティングシステムと、

前記コンピューティングシステムと動作可能に通信する複数のカメラセンサと

を備え、

前記コンピューティングシステムは、

前記カメラセンサからの前記カメラセンサの視野内の場面のアンカ画像と、場面の基準画像のセットとを含むマルチビュー画像を取得することと、

(1) 前記アンカ画像および各基準画像に関する記述子の個別の記述子フィールドを出力し、(i i) 相対的姿勢と併せて、前記アンカ画像内の疎着目点を検出し、代替視点から前記基準画像内の検索空間を決定し、(i i i) 中間特徴マップを出力する、共有RGBエンコーダおよび記述子デコーダを通して、前記アンカ画像および基準画像を通過させることであって、前記記述子デコーダは、記述子を訓練するように構成される学習アーキテクチャを利用する、ことと、

各基準画像の前記検索空間内の前記個別の記述子をサンプリングし、前記検索空間内

10

20

30

40

50

の記述子を決定し、前記決定された記述子と前記アンカ画像内の前記疎着目点に関する記述子をマッチングさせることであって、そのようなマッチングされる記述子は、マッチングされた特徴点と称される、ことと、

特異値分解（SVD）を使用して、前記マッチングされた特徴点を三角測量し、3D点を出力することと、

前記3D点を疎深度エンコーダを通して通過させ、疎深度画像を前記3D点および出力特徴マップから作成することと、

深度デコーダが、前記疎深度エンコーダに関する前記出力特徴マップと、前記共有RGBエンコーダからの前記中間特徴マップとに基づいて、稠密深度画像を生成することを含むプロセスによって、前記カメラセンサによって捕捉された複数のマルチビュー画像からの場面内の特徴の深度を推定するように構成される、クロスリアリティシステム。

【請求項9】

前記共有RGBエンコーダおよび記述子デコーダは、前記アンカ画像およびトランザクション画像の完全分解能に作用するように構成される完全畳み込みニューラルネットワークである、請求項8に記載のクロスリアリティシステム。

【請求項10】

前記カメラセンサによって捕捉された複数のマルチビュー画像からの場面内の特徴の深度を推定するためのプロセスは、

前記共有RGBエンコーダからの前記中間特徴マップを第1のタスク特有のデコーダヘッドの中にフィードし、前記アンカ画像内の前記疎着目点の検出のための加重を決定し、着目点記述を出力すること

をさらに含む、請求項8に記載のクロスリアリティシステム。

【請求項11】

前記記述子デコーダは、U-Net様アーキテクチャを備え、前記決定された記述子と前記疎着目点に関する記述子をマッチングさせるために、微細および粗大レベル画像情報を融合させる、請求項8に記載のクロスリアリティシステム。

【請求項12】

前記検索空間は、前記基準画像内の個別のエピポーラ線および前記エピポーラ線の両側上の固定されたオフセットに制約され、前記エピポーラ線に沿った実行可能深度感知範囲内にある、請求項8に記載のクロスリアリティシステム。

【請求項13】

双線形サンプリングが、前記共有RGBエンコーダおよび記述子デコーダによって使用され、前記個別の記述子を前記記述子フィールド内の所望の点において出力する、請求項8に記載のクロスリアリティシステム。

【請求項14】

前記マッチングされた特徴点を三角測量するステップは、

空間軸を横断して、softmaxを算出し、相互相関マップを出力することによって、前記疎着目点の個別の2次元（2D）位置を推定することと、

soft-argmax演算を実施し、節点の2D位置を対応する相互相関マップの質量中心として計算することと、

線形代数三角測量を前記推定された個別の2D位置から実施することと、

特異値分解（SVD）を使用して、3D点を出力することと

を含む、請求項8に記載のクロスリアリティシステム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、コンピューティング、学習ネットワーク構成、および接続されたモバイルコンピューティングシステム、方法、および構成に関し、より具体的には、推定される深度が、仮想および／または拡張現実動作のために構成される、少なくとも1つのウェアラブルコンポーネントを特徴とする、モバイルコンピューティングシステム、方法、および構

10

20

30

40

50

成において使用され得る、マルチビュー画像からの場面内の特徴の深度を推定するためのシステムおよび方に関する。

【背景技術】

【0002】

現代のコンピューティングおよびディスプレイ技術は、集合的に、「クロスリアリティ」環境または体験と称される、いわゆる仮想現実（「VR」）、拡張現実（「AR」）、および／または複合現実（「MR」）体験のためのシステムの開発を促進している。これは、頭部搭載型ディスプレイを通して、コンピュータ生成画像をユーザに提示することによって行われることができる。本画像は、感覚体験を作成し、これは、ユーザをシミュレートされた環境に没入させる。本データは、例えば、ユーザが、物理的世界の一部として、感知または知覚し、仮想オブジェクトと相互作用し得るようにレンダリングされ得る、仮想オブジェクトを説明し得る。ユーザは、例えば、頭部搭載型ディスプレイデバイス等のユーザインターフェースデバイスを通してレンダリングおよび提示されているデータの結果として、これらの仮想オブジェクトを体験し得る。データは、ユーザに見えるように表示され得る、またはユーザに聞こえるように再生される、オーディオを制御し得る、または触知的（または触覚的）インターフェースを制御し、ユーザが、仮想オブジェクトを感じるにつれて、ユーザが感知または知覚する、タッチ感覚を体験することを可能にし得る。

10

【0003】

XRシステムは、科学的可視化、医療訓練、工学設計、およびプロトタイプ化、遠隔操作およびテレプレゼンス、および個人的娯楽の分野に及ぶ、多くの用途のために有用であり得る。VRシステムは、典型的には、実際の実世界視覚的入力に対する透過性を伴わずに、デジタルまたは仮想画像情報の提示を伴う。

20

【0004】

ARシステムは、概して、実世界環境をシミュレートされた要素で補完する。例えば、ARシステムは、ユーザに、頭部搭載型ディスプレイを介して、周囲の実世界環境のビューを提供し得る。コンピュータ生成画像はまた、周囲の実世界環境を向上させるために頭部搭載型ディスプレイ上に提示されることもできる。本コンピュータ生成画像は、周囲の実世界環境にコンテキスト的に関連する、要素を含むことができる。そのような要素は、シミュレートされたテキスト、画像、オブジェクト、および同等物を含むことができる。MRシステムはまた、シミュレートされたオブジェクトを実世界環境に導入するが、これらのオブジェクトは、典型的には、ARシステムを上回る相互作用の程度を特徴とする。

30

【0005】

AR／MRシナリオは、多くの場合、実世界オブジェクトに関連して仮想画像要素の提示を含む。例えば、AR／MR場面は、AR／MR技術のユーザに、構造、オブジェクト等を含む、ユーザを囲繞する環境を特徴とする、実世界場面が見えるように描写される。これらの特徴に加え、AR／MR技術のユーザは、そのような特徴が実世界環境内に存在しない場合でも、彼らにコンピュータ生成された特徴（すなわち、仮想オブジェクト）が「見える」と知覚する。故に、ARおよびMRは、VRと対照的に、物理的世界の実オブジェクトに関連して、1つまたはそれを上回る仮想オブジェクトを含む。仮想オブジェクトはまた、実世界オブジェクトと相互作用し、したがって、AR／MRシステムは、ユーザを囲繞する3D世界とのシステムの相互作用に関連して、「空間コンピューティング」システムとも称され得る。実オブジェクトと相互作用する仮想オブジェクトの体験は、XRシステムを使用する際、ユーザの享受を大幅に向上させ、また、物理的世界が改変され得る方法についての現実的かつ容易に理解可能な情報を提示する、種々の用途のための可能性を広げる。

40

【0006】

脳の視覚中枢は、貴重な知覚情報を相互に対する両眼およびその構成要素の運動から得る。相互に対する2つの眼の輻輳・開散運動移動（すなわち、眼の視線をオブジェクト上に収束させ、それを固視するための相互に向かってまたはそこから離れる瞳孔の回転移動

50

）は、眼の水晶体の遠近調節（または合焦）と緊密に関連付けられる。正常条件下、眼を遠近調節し、または眼の水晶体の焦点を変化させ、異なる距離におけるオブジェクト上に合焦させることは、「遠近調節－輻輳・開散運動反射」として知られる関係下、同一距離への輻輳・開散運動の整合変化を自動的に生じさせるであろう。同様に、輻輳・開散運動の変化が、正常条件下、遠近調節の合致する変化を誘起するであろう。本反射に逆らう作用は、大部分の従来の立体視VR／AR／MR構成におけるように、眼精疲労、頭痛、または他の形態の不快感をユーザにもたらすことが知られている。

【0007】

立体視ウェアラブル眼鏡は、概して、3次元視点がヒト視覚系によって知覚されるように、若干異なる要素提示を伴う画像を表示するように構成される、2つのディスプレイ（左眼に1つ、右眼に1つ）を特徴とする。そのような構成は、輻輳・開散運動と遠近調節との間の不整合（「輻輳・開散運動－遠近調節衝突」）に起因して、多くのユーザにとって不快であることが見出されており、これは、3次元における画像を知覚するために克服されなければならない。実際、一部のユーザは、立体視構成に耐えることが不可能である。これらの限界は、VR、AR、およびMRシステムに該当する。故に、大部分の従来のVR／AR／MRシステムは、部分的に、従来のシステムが、輻輳・開散運動－遠近調節衝突を含む、ヒト知覚系の基本側面のうちのいくつかに対処することができないため、ユーザにとって快適かつ最大限に有用となるであろう様式において、豊かな両眼3次元体験を提示するために最適に好適ではない。

【0008】

種々のシステムおよび方法が、輻輳・開散運動－遠近調節衝突に対処するために開示されている。例えば、米国特許出願第14／555,585号は、光および光によってレンダリングされた画像が複数の深度面から生じるように現れるように、1つまたはそれを上回る光誘導光学要素を使用して、光をユーザの眼に投影させることによって、輻輳・開散運動－遠近調節衝突に対処する、VR／AR／MRシステムおよび方法を開示する。光誘導光学要素は、デジタルまたは仮想オブジェクトに対応する仮想光を内部結合し、それを全内部反射（「TIR」）によって伝搬し、次いで、仮想光を外部結合し、仮想オブジェクトをユーザの眼に表示するように設計される。AR／MRシステムでは、光誘導光学要素はまた、実際の実世界オブジェクトからの（例えば、そこから反射する）光に対して透過性であるように設計される。したがって、光誘導光学要素の一部は、TIRを介した伝搬のために仮想光を反射させながら、AR／MRシステム内の実世界オブジェクトからの実世界光に対して透過性であるように設計される。

【0009】

AR／MRシナリオは、多くの場合、仮想オブジェクトと実世界の物理的環境との間の相互作用を含む。同様に、いくつかのVRシナリオは、完全仮想オブジェクトと他の仮想オブジェクトとの間の相互作用を含む。物理的環境内のオブジェクトの輪郭を描くことは、それらの相互作用の境および境界線を画定することによって（例えば、物理的環境内の特定の構造またはオブジェクトの範囲を画定することによって）、仮想オブジェクトとの相互作用を促進する。例えば、AR／MRシナリオが、物理的環境内の特定のオブジェクトから延在する仮想オブジェクト（例えば、触手または拳）を含む場合、3次元内のオブジェクトの範囲を画定することは、AR／MRシステムがより現実的AR／MRシナリオを提示することを可能にする。逆に言えば、オブジェクトの範囲が、画定されない、または不正確に画定される場合、アーチファクトおよび誤差が、表示される画像に生じるであろう。例えば、仮想オブジェクトは、オブジェクトの表面からの代わりに、部分的または完全に、オブジェクトに隣接する空中から延在するように現れ得る。別の実施例では、AR／MRシナリオが、物理的環境内の特定の水平表面上を歩いている仮想キャラクタを含む場合、表面の範囲を不正確に画定することは、落下せずに、代わりに、空中に浮遊して、表面から離れて歩いているように現れる、仮想キャラクタをもたらし得る。

【0010】

故に、周囲環境等の場面の深度感知は、クロスリアリティシステムから自律運転に及ぶ

10

20

30

40

50

、広範囲の用途のために有用である。場面の深度を推定することは、広義には、以下のクラス、すなわち、能動的小よび受動的感知に分割されることができろ。能動的感知技法は、L i D A R、構造化光、および飛行時間 (T o F) カメラを含む一方、単眼カメラまたはカメラのアレイの立体視を使用する、深度推定は、受動的感知と称される。能動的センサは、現在、事実上、可変環境内における良好な正確度および短待ち時間に起因して、深度感知を要求する用途の標準である ([参考文献 4 4] 参照)。括弧内の付番された参考文献 (「[参考文献 # #]」) は、下記に添付の参考文献一覧を指し、これらの参考文献はそれぞれ、参照することによってその全体として本明細書に組み込まれる。

【0011】

しかしながら、能動的センサは、その独自の限界を有する。L i D A Rは、法外に高価であって、かつ疎測定を提供する。構造化光およびT o F深度カメラは、光輸送の物理学に起因して、限定された範囲および完全性を有する。さらに、それらは、電力を大量消費し、ウェアラブル上のAR／VR用途のために重要なモビリティを阻害する。その結果、コンピュータビジョン研究者は、能動的センサの普遍的、コスト効果的、かつエネルギー効率的代替として、受動的感知技法を追求している ([参考文献 3 0] 参照)。

【0012】

ステレオカメラを使用する、受動的深度感知は、正確な深度推定のために大規模ベースラインおよび慎重な較正を要求する ([参考文献 3] 参照)。大規模ベースラインは、電話およびウェアラブルのようなモバイルデバイスにとって、実行不可能である。代替は、可動単眼カメラのためのマルチビューステレオ (M V S) 技法を使用して、深度を推定するものである。M V Sは、概して、既知のカメラ姿勢および固有性質を伴う、複数の画像から、3 D場面構造を再構築する問題を指す ([参考文献 1 4] 参照)。カメラ運動の制約されない本質は、ステレオリグのベースライン限界を緩和し、アルゴリズムは、持続的に変動する視点からの同一場面の複数の観察から利益を享受する ([参考文献 1 7] 参照)。しかしながら、カメラ運動はまた、姿勢不確実性および運動アーチファクトの追加される複雑性に起因して、剛性ステレオリグに対して、深度推定をより困難にする。大部分のM V Sアプローチは、通常、平面掃引ステレオアプローチを用いて、3 Dコスト体積を構築することを伴う ([参考文献 4 1、1 8] 参照)。M V Sを使用する、正確な深度推定は、コスト体積上での3 D畳み込みに依拠し、これは、メモリ上および算出上の両方において高価であって、分解能の3乗に比例する。さらに、冗長算出が、カメラ姿勢推定、故に、任意のM V S技法にとって必要な先行処理である、着目点およびその記述子等の有用な画像レベル性質を無視することによって、追加される。これは、受動的感知のための全体的コストおよびエネルギー要件を増加させる。

【0013】

単一画像を使用する、受動的感知は、基本的に、2 D画像内のスケール曖昧性に起因して、信頼性がない。深層学習ベースの単眼深度推定アプローチは、問題を深度回帰として公式化し ([参考文献 1 0、1 1] 参照)、能動的センサのものとの性能上の間隙を低減させている ([参考文献 2 6、2 4] 参照) が、依然として、実践的のものには遠く及ばない。最近、疎／稠密深度推定アプローチが、スケール曖昧性を除去し、単眼深度推定のロバスト性を改良するために提案されている ([参考文献 3 0] 参照)。実際、0.5%未満の深度サンプルを伴う、最近の疎／稠密アプローチは、より広い範囲および高い完全性を伴って、能動的センサと同等である正確度を有する、([参考文献 6] 参照)。しかしながら、これらのアプローチは、能動的センサからの正確なまたはシード深度サンプルを仮定し、これは、限定的である。代替は、同時位置特定およびマッピング (S L A M) ([参考文献 3 1] 参照) または視覚慣性オドメトリ (V I O) ([参考文献 3 3] 参照) のための最良性能アルゴリズムから出力された疎3 D目印を使用するものである。しかしながら、能動的センサからの深度の代わりに、これらの疎目印から評価される深度を使用することは、性能を有意に劣化させる ([参考文献 4 3] 参照)。これは、学習された疎／稠密ネットワークが、S L A MまたはV I Oアルゴリズム内に存在する潜在的に有用なキュー、構造化雑音、およびバイアスを無視するため、驚くべきことではない。

10

20

30

40

50

【0014】

疎特徴ベースの方法は、その高速性および正確度に起因して、SLAMまたはVIO技法のための標準である。検出後記述アプローチは、疎特徴抽出に対する最も一般的アプローチであって、着目点が、検出され、次いで、点の周囲のパッチに関して記述される。記述子は、角、プロブ等の典型的低レベル着目点によって欠失される、より高いレベルの情報をカプセル化する。深層学習革命以前は、SIFT（[参考文献28]参照）およびORB（[参考文献37]参照）のような古典的システムが、低レベル視覚タスクに関する特徴マッチングのための記述子として普遍的に使用されていた。目前の目的を直接最適化する、深層ニューラルネットワークは、現在、多様な用途を横断して、これらの手動で作られた特徴に取って代わっている。しかしながら、そのようなエンドツーエンドネットワークは、成分が非可微分であることに起因して、SLAM（[参考文献32]参照）にとって分かりにくいままである。SuperPoint（[参考文献9]参照）、LIFT（[参考文献42]参照）、およびGIFT（[参考文献27]参照）等の方法によって学習された汎用記述子は、可微分SLAMに向かって、間隙を埋めることを狙いとする。

【0015】

MVSアプローチは、3D体積を直接再構築するか、または3D再構築物または他の用途のためにフレキシブルに使用され得る、深度マップを出力するかのいずれかである。3D体積を再構築する方法（[参考文献41、5]参照）は、3Dボクセル化空間で動作する高メモリ負荷に起因して（[参考文献35、39]参照）、または複雑な環境内での点表現を学習することの困難度に起因して（[参考文献34]参照）のいずれかにおいて、小空間または隔離されたオブジェクトに制限される。屋内環境内で捕捉されたマルチビュー画像の使用が、最近、学習されたパッチマッチングアプローチを提案した、DeepMVS（[参考文献18]参照）から開始して、進歩している。MVDepthNet（[参考文献40]参照）およびDPSNet（[参考文献19]参照）は、深度推定のためのコスト体積を構築する。最近、GP-MVSNet（[参考文献17]参照）が、MVDepthNetを踏まえ、ガウスプロセスを使用して、時間的情報をコヒーレントに融合させた。全てのこれらの方法は、深度推定のある段階の間、平面掃引アルゴリズムを利用し、正確度対効率性のトレードオフをもたらす。

【0016】

疎／稠密深度推定もまた、最近、電力予算に基づいて動作するときのその範囲限界に起因して、能動的深度センサを補完するため、および暗いまたは反射性オブジェクト等の検出が困難な領域における深度を充填するための方法として、出現している。1つのアプローチが、Ma et al（[参考文献30]参照）によって提案され、その後、Chen et al（[参考文献6、43]参照）が続き、これは、革新を表現およびネットワークアーキテクチャに導入した。畳み込み空間伝搬モジュールが、欠失深度値を充填するために提案される[参考文献7]。最近、自己教師ありアプローチ（[参考文献13、12]参照）が、疎／稠密問題に関して模索されている（[参考文献29]参照）。

【0017】

マルチビューステレオ（MVS）は、能動的深度感知の正確度と単眼深度推定の実践性との間の有利な中間アプローチを表すことが分かる。3D畳み込みニューラルネットワーク（CNN）を採用する、コスト体積ベースのアプローチは、MVSシステムの正確度を著しく改良している。しかしながら、本正確度は、高算出コストの結果となり、これは、実践的採用を妨害する。

【0018】

故に、コストがかかり、かつ非効果的能動的深度感知に依存せず、かつ以前の受動的深度感知技法の効率および／または正確度を改良する、場面の深度推定のための改良されたシステムおよび方法の必要性が存在する。加えて、軽量、低コスト、小形状因子を有し、広仮想画像視野を有し、可能な限り透明である、ディスプレイを有する、深度推定のためのシステムおよび方法は、XRシステム内に実装可能であるべきである。

【発明の概要】

【課題を解決するための手段】

【0019】

本明細書に開示される実施形態は、エンドツーエンドプロセスにおいて、XRシステム等の空間コンピューティングシステムのユーザを囲繞する、場面または環境内の特徴の深度を推定するためのシステムおよび方法を対象とする。推定される深度は、空間コンピューティングシステムによって、例えば、正確かつ効果的3DXR体験を提供するために利用されることができる。結果として生じる3DXR体験は、部分的に、輻輳・開散運動－遠近調節不整合等のヒト知覚系の基本側面のうちのいくつかに対処する様式において、画像を提示し得るため、ユーザにとって快適かつ最大限に有用である、豊かな両眼3次元体験において表示可能である。例えば、推定される深度は、正確な深度データを有し、3D画像が複数の焦点面内に表示されることを有効にする、3D再構築物を生成するために使用されてもよい。3D再構築物はまた、仮想オブジェクト、他の仮想オブジェクト、および／または実世界オブジェクト間の相互作用の正確な管理を有効にしてもよい。

10

【0020】

故に、一実施形態は、マルチビュー画像からの場面内の特徴の深度を推定するための方法を対象とする。最初に、場面のアンカ画像と、場面の基準画像のセットとを含む、マルチビュー画像が、取得される。これは、XRシステムのカメラ等の1つまたはそれを上回る好適なカメラによって遂行され得る。アンカ画像および基準画像は、(1)アンカ画像および各基準画像に関する記述子の個別の記述子フィールドを出力し、(ii)相対的姿勢と併せて、アンカ画像内の着目点を検出し、代替視点から基準画像内の検索空間を決定し、(iii)中間特徴マップを出力する、共有RGBエンコーダおよび記述子デコーダを通して通過される。個別の記述子は、各基準画像の検索空間内でサンプリングされ、検索空間内の記述子を決定し、識別された記述子とアンカ画像内の着目点に関する記述子をマッチングさせる。マッチングされる記述子は、マッチングされた特徴点と称される。マッチングされた特徴点は、特異値分解(SVD)を使用して三角測量され、3D点を出力する。3D点は、疎深度エンコーダを通して通過され、疎深度画像を3D点から作成し、特徴マップを出力する。深度デコーダは、次いで、疎深度エンコーダに関する出力特徴マップと、RGBエンコーダからの中間特徴マップとに基づいて、稠密深度画像を生成する。

20

【0021】

本方法の別の側面では、共有RGBエンコーダおよび記述子デコーダは、RGB画像エンコーダおよび疎深度画像エンコーダを含む、2つのエンコーダと、着目点検出エンコーダ、記述子デコーダ、および稠密深度予測エンコーダを含む、3つのデコーダとを備えてもよい。

30

【0022】

さらに別の側面では、共有RGBエンコーダおよび記述子デコーダは、アンカ画像およびトランザクション画像の完全分解能に作用するように構成される、完全畳み込みニューラルネットワークであってもよい。

【0023】

さらに別の側面では、本方法はさらに、RGBエンコーダからの特徴マップを第1のタスク特有のデコーダヘッドの中にフィードし、アンカ画像内の着目点の検出のための加重を決定し、着目点記述を出力するステップを含んでもよい。

40

【0024】

本方法のさらに別の側面では、記述子デコーダは、U-Net様アーキテクチャを備え、識別された記述子と着目点に関する記述子をマッチングさせるための微細および粗大レベル画像情報を融合させてもよい。

【0025】

本方法の別の側面では、検索空間は、基準画像内の個別のエピポーラ線＋エピポーラ線の両側上の固定されたオフセットに制約され、エピポーラ線に沿った実行可能深度感知範囲内にあってもよい。

【0026】

50

さらに本方法の別の側面では、双線形サンプリングが、共有RGBエンコーダおよび記述子デコーダによって使用され、個別の記述子を記述子フィールド内の所望の点において出力してもよい。

【0027】

本方法の別の側面では、マッチングされた特徴点を三角測量するステップは、空間軸を横断して、 softmax を算出し、相互相関マップを出力することによって、着目点の個別の2次元(2D)位置を推定するステップと、 soft-argmax 演算を実施し、節点の2D位置を対応する相互相関マップの質量中心として計算するステップと、線形代数三角測量を2D推定値から実施するステップと、特異値分解(SVD)を使用して、3D点を出力するステップとを含む。

【0028】

別の開示される実施形態は、クロスリアリティ(XR)システムを対象とし、これは、深度を推定するように構成され、本明細書に説明されるような深度のために利用される。クロスリアリティシステムは、ディスプレイシステムを有する、頭部搭載型ディスプレイデバイスを備える。例えば、頭部搭載型ディスプレイは、一对の接眼ディスプレイを眼鏡状構造内に有してもよい。コンピューティングシステムは、頭部搭載型ディスプレイと動作可能に通信する。複数のカメラセンサは、コンピューティングシステムと動作可能に通信する。コンピューティングシステムは、上記に説明される方法のいずれかにおいて、カメラセンサによって捕捉された複数のマルチビュー画像からの場面内の特徴の深度を推定するように構成される。クロスリアリティシステムの付加的側面では、プロセスは、上記に説明されるクロスリアリティシステムの付加的側面のうちの任意の1つまたはそれを上回るものを含んでもよい。例えば、プロセスは、カメラセンサからのカメラセンサの視野内の場面のアンカ画像と、場面の基準画像のセットとを含む、マルチビュー画像を取得するステップと、(1)アンカ画像および各基準画像に関する記述子の個別の記述子フィールドを出力し、(ii)相対的姿勢と併せて、アンカ画像内の着目点を検出し、代替視点から基準画像内の検索空間を決定し、(iii)中間特徴マップを出力する、共有RGBエンコーダおよび記述子デコーダを通して、アンカ画像および基準画像を通過させるステップと、各基準画像の検索空間内の個別の記述子をサンプリングし、検索空間内の記述子を決定し、識別された記述子とアンカ画像内の着目点に関する記述子をマッチングさせるステップであって、そのようなマッチングされる記述子は、マッチングされた特徴点と称される、ステップと、特異値分解(SVD)を使用して、マッチングされた特徴点を三角測量し、3D点を出力するステップと、3D点を疎深度エンコーダを通して通過させ、疎深度画像を3D点および出力特徴マップから作成するステップと、深度デコーダが、疎深度エンコーダに関する出力特徴マップと、RGBエンコーダからの中間特徴マップとに基づいて、稠密深度画像を生成するステップとを含んでもよい。

本発明は、例えば、以下を提供する。

(項目1)

マルチビュー画像からの場面内の特徴の深度を推定するための方法であって、前記方法は、

前記場面のアンカ画像と、前記場面の基準画像のセットとを含むマルチビュー画像を取得するステップと、

(1)前記アンカ画像および各基準画像に関する記述子の個別の記述子フィールドを出力し、(ii)相対的姿勢と併せて、前記アンカ画像内の着目点を検出し、代替視点から前記基準画像内の検索空間を決定し、(iii)中間特徴マップを出力する、共有RGBエンコーダおよび記述子デコーダを通して、前記アンカ画像および基準画像を通過させるステップと、

各基準画像の前記検索空間内の前記個別の記述子をサンプリングし、前記検索空間内の記述子を決定し、前記識別された記述子と前記アンカ画像内の着目点に関する記述子をマッチングさせるステップであって、そのようなマッチングされる記述子は、マッチングされた特徴点と称される、ステップと、

10

20

30

40

50

特異値分解（SVD）を使用して、前記マッチングされた特徴点を三角測量し、3D点
を出力するステップと、

前記3D点を疎深度エンコーダを通して通過させ、疎深度画像を前記3D点および出力
特徴マップから作成するステップと、

深度デコーダが、前記疎深度エンコーダに関する前記出力特徴マップと、前記RGBエ
ンコーダからの前記中間特徴マップとに基づいて、稠密深度画像を生成するステップと
を含む、方法。

（項目2）

前記共有RGBエンコーダおよび記述子デコーダは、RGB画像エンコーダおよび疎深
度画像エンコーダを含む2つのエンコーダと、着目点検出エンコーダ、記述子デコーダ、
および稠密深度予測エンコーダを含む3つのデコーダとを備える、項目1に記載の方法。

（項目3）

前記共有RGBエンコーダおよび記述子デコーダは、前記アンカ画像およびトランザク
ション画像の完全分解能に作用するように構成される完全畳み込みニューラルネットワ
ークである、項目1に記載の方法。

（項目4）

前記RGBエンコーダからの前記特徴マップを第1のタスク特有のデコーダヘッドの中
にフィードし、前記アンカ画像内の着目点の検出のための加重を決定し、着目点記述を出力
するステップ

をさらに含む、項目1に記載の方法。

（項目5）

前記記述子デコーダは、U-Net様アーキテクチャを備え、前記識別された記述子と
前記着目点に関する記述子をマッチングさせるために、微細および粗大レベル画像情報を
融合させる、項目1に記載の方法。

（項目6）

前記検索空間は、前記基準画像内の個別のエピポーラ線+前記エピポーラ線の両側上の
固定されたオフセットに制約され、前記エピポーラ線に沿った実行可能深度感知範囲内に
ある、項目1に記載の方法。

（項目7）

双線形サンプリングが、前記共有RGBエンコーダおよび記述子デコーダによって使用
され、前記個別の記述子を前記記述子フィールド内の所望の点において出力する、項目1
に記載の方法。

（項目8）

前記マッチングされた特徴点を三角測量するステップは、
空間軸を横断して、softmaxを算出し、相互相関マップを出力することによって
、前記着目点の個別の2次元（2D）位置を推定するステップと、

softmax演算を実施し、節点の2D位置を対応する相互相関マップの質
量中心として計算するステップと、

線形代数三角測量を前記2D推定値から実施するステップと、

特異値分解（SVD）を使用して、3D点を出力するステップと

を含む、項目1に記載の方法。

（項目9）

クロスリアリティシステムであって、

ディスプレイシステムを有する頭部搭載型ディスプレイデバイスと、

前記頭部搭載型ディスプレイと動作可能に通信するコンピューティングシステムと、

前記コンピューティングシステムと動作可能に通信する複数のカメラセンサと

を備え、前記コンピューティングシステムは、

前記カメラセンサからの前記カメラセンサの視野内の場面のアンカ画像と、場面の基準
画像のセットとを含むマルチビュー画像を取得するステップと、

（1）前記アンカ画像および各基準画像に関する記述子の個別の記述子フィールドを出

10

20

30

40

50

かし、(i i) 相対的姿勢と併せて、前記アンカ画像内の着目点を検出し、代替視点から前記基準画像内の検索空間を決定し、(i i i) 中間特徴マップを出力する、共有RGBエンコーダおよび記述子デコーダを通して、前記アンカ画像および基準画像を通過させるステップと、

各基準画像の前記検索空間内の前記個別の記述子をサンプリングし、前記検索空間内の記述子を決定し、前記識別された記述子と前記アンカ画像内の着目点に関する記述子をマッチングさせるステップであって、そのようなマッチングされる記述子は、マッチングされた特徴点と称される、ステップと、

特異値分解(SVD)を使用して、前記マッチングされた特徴点を三角測量し、3D点を出力するステップと、

前記3D点を疎深度エンコーダを通して通過させ、疎深度画像を前記3D点および出力特徴マップから作成するステップと、

深度デコーダが、前記疎深度エンコーダに関する前記出力特徴マップと、前記RGBエンコーダからの前記中間特徴マップとに基づいて、稠密深度画像を生成するステップと

を含むプロセスによって、前記カメラセンサによって捕捉された複数のマルチビュー画像からの場面内の特徴の深度を推定するように構成される、クロスリアリティシステム。

(項目10)

前記共有RGBエンコーダおよび記述子デコーダは、RGB画像エンコーダおよび疎深度画像エンコーダを含む2つのエンコーダと、着目点検出エンコーダ、記述子デコーダ、および稠密深度予測エンコーダを含む3つのデコーダとを備える、項目9に記載のクロスリアリティシステム。

(項目11)

前記共有RGBエンコーダおよび記述子デコーダは、前記アンカ画像およびトランザクション画像の完全分解能に作用するように構成される完全畳み込みニューラルネットワークである、項目9に記載のクロスリアリティシステム。

(項目12)

前記カメラセンサによって捕捉された複数のマルチビュー画像からの場面内の特徴の深度を推定するためのプロセスはさらに、

前記RGBエンコーダからの前記特徴マップを第1のタスク特有のデコーダヘッドの中にフィードし、前記アンカ画像内の着目点の検出のための加重を決定し、着目点記述を出力するステップ

を含む、項目9に記載のクロスリアリティシステム。

(項目13)

前記記述子デコーダは、U-Net様アーキテクチャを備え、前記識別された記述子と前記着目点に関する記述子をマッチングさせるために、微細および粗大レベル画像情報を融合させる、項目9に記載のクロスリアリティシステム。

(項目14)

前記検索空間は、前記基準画像内の個別のエピポーラ線+前記エピポーラ線の両側上の固定されたオフセットに制約され、前記エピポーラ線に沿った実行可能深度感知範囲内にある、項目9に記載のクロスリアリティシステム。

(項目15)

双線形サンプリングが、前記共有RGBエンコーダおよび記述子デコーダによって使用され、前記個別の記述子を前記記述子フィールド内の所望の点において出力する、項目9に記載のクロスリアリティシステム。

(項目16)

前記マッチングされた特徴点を三角測量するステップは、空間軸を横断して、softmaxを算出し、相互相関マップを出力することによって、前記着目点の個別の2次元(2D)位置を推定するステップと、

softmaxargmax演算を実施し、節点の2D位置を対応する相互相関マップの質量中心として計算するステップと、

10

20

30

40

50

線形代数三角測量を前記 2 D 推定値から実施するステップと、
特異値分解 (SVD) を使用して、3 D 点を出力するステップと
を含む、項目 9 に記載のクロスリアリティシステム。

【図面の簡単な説明】

【0029】

図面は、本開示の好ましい実施形態の設計および有用性を図示し、その中で類似要素は、共通参照番号によって参照される。本開示の上記および他の利点および目的が取得される方法をより深く理解するために、上記に簡単に説明される、本開示のより具体的説明が、付随の図面に図示される、その具体的実施形態を参照することによって与えられるであろう。これらの図面は、本開示の典型的実施形態のみを描写し、したがって、その範囲の限定と見なされないことを理解した上で、本開示は、付随の図面の使用を通して、付加的

10

具体性および詳細とともに記載および説明されるであろう。

【0030】

【図1】図1は、一実施形態による、クロスリアリティ体験を提供するための例示的クロスリアリティシステムの概略図である。

【0031】

【図2】図2は、一実施形態による、場面の深度推定のための方法の概略図である。

【0032】

【図3】図3は、一実施形態による、図2の方法内で使用される、共有RGBエンコーダおよび記述子デコーダのアーキテクチャのブロック図である。

20

【0033】

【図4】図4は、一実施形態による、図2の方法内で使用されるように、エピソードサンプリングおよび深度範囲サンプリングを使用して、検索空間の範囲を制限するためのプロセスを図示する。

【0034】

【図5】図5は、一実施形態による、図2の方法内で使用されるように、特徴点ネットワークのためのアーキテクチャを図示する、ブロック図である。

【0035】

【図6】図6は、図2の方法の実施例と種々の他の異なる方法との間の定質的比較を図示する。

30

【0036】

【図7】図7は、本明細書に説明される、図2の方法の実施例の推定される深度マップIからの場面のサンプル3D再構築物を示す。

【0037】

【図8】図8は、ScanNet上での異なる記述子の性能の比較を有する、表1を示す。

【0038】

【図9】図9は、ScanNet上での深度推定の性能の比較を有する、表2を示す。

【0039】

【図10】図10は、異なる数の画像に関するScanNet上での深度推定の性能の比較を有する、表3を示す。

40

【0040】

【図11】図11は、Sun3D上での深度推定の比較を有する、表4を示す。

【0041】

【図12】図12は、一実施形態による、図2の方法内で使用されるように、画像視点毎に、その対応するエピソード線に沿って、記述子フィールドで畳み込まれる各着目点の記述子のためのプロセスに関する方程式を記載する。

【0042】

【図13】図13-16は、一実施形態による、図2の方法内で使用されるように、3D点を取得するための代数三角測量のためのプロセスに関する方程式を記載する。

【図14】図13-16は、一実施形態による、図2の方法内で使用されるように、3D

50

点を取得するための代数三角測量のためのプロセスに関する方程式を記載する。

【図 1 5】図 1 3 - 1 6 は、一実施形態による、図 2 の方法内で使用されるように、3 D 点を取得するための代数三角測量のためのプロセスに関する方程式を記載する。

【図 1 6】図 1 3 - 1 6 は、一実施形態による、図 2 の方法内で使用されるように、3 D 点を取得するための代数三角測量のためのプロセスに関する方程式を記載する。

【発明を実施するための形態】

【0 0 4 3】

詳細な説明

以下は、エンドツーエンドプロセスにおいて、X R システム等の空間コンピューティングシステムのユーザを囲繞する、場面または環境内の特徴の深度を推定するためのシステムおよび方法の種々の実施形態を説明する。種々の実施形態が、ここで、当業者が本開示を実践することを可能にするように、本開示の例証的实施例として提供される、図面を参照して詳細に説明されるであろう。着目すべきこととして、以下の図および実施例は、本開示の範囲を限定することを意味するものではない。本開示のある要素が、部分的または完全に、公知のコンポーネント（または方法またはプロセス）を使用して実装され得る場合、本開示の理解のために必要なそのような公知のコンポーネント（または方法またはプロセス）の一部のみが、説明され、そのような公知のコンポーネント（または方法またはプロセス）の他の部分の詳細な説明は、本開示を曖昧にしないように、省略されるであろう。さらに、種々の実施形態は、本明細書に例証として参照されるコンポーネントの現在および将来的公知の均等物を包含する。

【0 0 4 4】

さらに、空間コンピューティングシステムのユーザを囲繞する、場面または環境内の特徴の深度を推定するためのシステムおよび方法はまた、X R システムから独立して実装されてもよく、本明細書に描写される実施形態は、例証的目的のみのために、X R システムに関連して説明される。

【0 0 4 5】

図 1 を参照すると、一実施形態による、例示的 X R システム 1 0 0 が、図示される。X R システム 1 0 0 は、頭部搭載型ディスプレイデバイス 2（ヘッド装着型視認コンポーネント 2 ととも称される）と、ハンドヘルドコントローラ 4（ハンドヘルドコントローラコンポーネント 4 ととも称される）と、ベルトパックまたは同等物として、ユーザ上に装着されるように構成され得る、相互接続された補助コンピューティングシステムまたはコントローラ 6（相互接続される補助コンピューティングシステムまたはコントローラコンポーネント 6 ととも称される）とを含む。これらのコンポーネントはそれぞれ、IEEE 8 0 2 . 1 1、Bluetooth（登録商標）（R T M）、および他のコネクティビティ規格および構成によって規定されたもの等の有線または無線通信接続 1 0、1 2、1 4、1 6、1 7、1 8 を介して、相互および他の接続されたリソース 8（クラウドコンピューティングまたはクラウド記憶リソース等）と動作可能に通信する（すなわち、動作可能に結合される）。頭部搭載型ディスプレイデバイス 2 は、それを通してユーザには、その周囲の世界が、関連付けられるシステムコンポーネントによって生産されたビデオ画像および視覚的コンポーネントとともに見え得る、2 つの描写される光学要素 2 0 を含み、一対の画像源（例えば、マイクロディスプレイパネル）と、拡張現実体験のためにコンピュータ生成された画像を光学要素 2 0 上に表示するための視認光学系とを含む。図示される実施形態では、頭部搭載型ディスプレイデバイス 2 および対の画像源は、軽量、低コスト、小形状因子を有し、広仮想画像視野を有し、かつ可能な限り透明である。図 1 に図示されるように、X R システム 1 0 0 はまた、限定ではないが、種々のカメラタイプセンサ 2 2、2 4、2 6（モノクロ、カラー／R G B、および／または熱等）、深度カメラセンサ 2 8、および／または音センサ 3 0（マイクロホン等）を含む、ユーザの環境周囲に関する情報を提供するように構成される、種々のセンサを含む。

【0 0 4 6】

加えて、X R システム 1 0 0 は、輻輳・開散運動－遠近調節不整合に関する容認可能許

容値を超えずに、様々なユースケースのために実践的であるために、仮想画像情報を複数の焦点面（例えば、2つまたはそれを上回る）内に提示するように構成されることが望ましい。米国特許出願第14/555,585号、第14/690,401号、第14/331,218号、第15/481,255号、第62/627,155号、第62/518,539号、第16/229,532号、第16/155,564号、第15/413,284号、第16/020,541号、第62,702,322号、第62/206,765号、第15,597,694号、第16/221,065号、第15/968,673号、および第62/682,788号（それぞれ、参照することによってその全体として本明細書に組み込まれる）は、XRシステム100およびそのコンポーネントの種々の側面をさらに詳細に説明する。

【0047】

種々の実施形態では、ユーザが、図1に描写されるXRシステム100等の拡張現実システムを装着し、これは、動作されるときユーザの周囲の3次元世界とのそのようなシステムの相互作用に関連して、「空間コンピューティング」システムとも称され得る。カメラ22、24、26およびコンピューティングシステム6が、ユーザの環境周囲をマッピングし、および／または壁、床、椅子、および同等物等のユーザの環境周囲内の種々のオブジェクトの幾何学形状を表す種々の点を備える、そのような環境の「メッシュ」を作成するように構成される。空間コンピューティングシステムは、ユーザの周囲の環境をマッピングまたはメッシュ化し、Magic Leap, Inc. (Planation, Florida) から利用可能なもの等のソフトウェアを起動または動作させるように構成されてもよく、これは、部屋のマップまたはメッシュを利用して、ユーザが、ユーザの周囲の3次元空間内の種々のオブジェクトおよび要素を設置、操作、可視化、作成、および修正することを補助するように構成されてもよい。図1に示されるように、XRシステム100はまた、クラウドまたは他のコンネクティビティ構成によって、他のコンピューティングシステム等の付加的接続されたリソース8に動作可能に結合されてもよい。

【0048】

本明細書に説明される方法、システム、および構成は、適切なセンサおよび利用可能な関連付けられるデータに従って、XRシステム100等のウェアラブル空間コンピューティングの領域の外側の種々のシナリオに広く適用可能であることを理解されたい。

【0049】

場面の深度推定のための以前のシステムおよび方法と対照的に、本開示のシステムおよび方法は、(a) MVS技法におけるコスト体積への依存性を除去し、したがって、算出量を有意に低減させ、(b) 検出された着目点および記述子を再使用することによって、疎VIOまたはSLAMを使用して、カメラ姿勢推定を補完し、(c) 幾何学形状ベースのMVS概念を利用して、アルゴリズムを誘導し、解釈可能性を改良し、(d) 疎／稠密技法の正確度および効率性から利益を享受するように、エンドツーエンド様式において、疎／稠密公式と併せて、疎3D目印を学習する。本システムおよび方法におけるネットワークは、RGB画像のために1つと、疎深度画像のために1つとの、2つのエンコーダと、着目点検出のために1つと、記述子のために1つと、稠密深度予測のために1つとの、3つのデコーダとから成る、エンコーダーデコーダ構造で構成される、マルチタスクモデル（[参考文献22]参照）である。効率的に、幾何学的初期値を使用して、点を三角測量し、着目点デコーダと、記述子デコーダと、疎深度エンコーダとの間の重要な連結を形成し、エンドツーエンド訓練を有効にする、可微分モジュールもまた、利用される。

【0050】

これらの方法および構成は、適切なセンサおよび関連付けられるデータが利用可能であることを条件として、ウェアラブル空間コンピューティングの領域外の種々のシナリオにも広く適用可能である。

【0051】

空間コンピューティングにおける課題のうちの1つは、例えば、ユーザの周囲の3次元世界に関連し得る、コンピュータビジョンおよび／またはオブジェクト認識課題等におい

10

20

30

40

50

て、ユーザに有用および／または重要な決定を行う際に、XRシステム100の種々の動作可能に結合されるセンサ（図1のシステム100の要素22、24、26、28等）によって捕捉されたデータの利用に関する。本明細書に開示されるものは、深度センサ28からの深度データを使用せずに、カメラ22、24、および26からのRGB画像等のRGB画像のみを使用して、XRシステム100のユーザを囲繞する3D環境等の場面の3D再構築物を生成するための方法およびシステムである。

【0052】

屋内環境等の場面の深度推定の以前の方法と対照的に、本開示は、マルチビューステレオのための疎点の三角測量および緻密化を学習することによって、深度推定のためのアプローチを導入する。コスト体積アプローチと明確に異なり、本開示のシステムおよび方法は、最初に、（a）着目点を検出し、それに関する記述子を評価し、次いで、（b）着目点の小集合をマッチングおよび三角測量するように学習し、最終的に、CNNを使用して、3D点の本疎集合を稠密化することによって、効率的深度推定アプローチを利用する。エンドツーエンドネットワークは、効率的に、全3つのステップを深層学習フレームワーク内で実施し、深度監視とともに、中間2D画像および3D幾何学的監視を用いて、訓練される。重大なこととして、本開示の方法の第1のステップは、着目点検出および記述子学習を使用して、姿勢推定を補完する。本方法は、より低い算出量を伴って、異なる場面長に関して、最先端結果を深度推定上で生産することが示される。さらに、本方法は、より新しい環境に対して一般化され、ネットワークによって出力された記述子は、優先的に、強固なベースラインと比較される。

【0053】

本開示の方法では、疎3D目印は、（a）MVS技法におけるようなコスト体積への依存性を除去し、したがって、算出コストを有意に低減させ、（b）検出された着目点および記述子を再使用することによって、疎VIOまたはSLAMを使用して、カメラ姿勢推定を補完し、（c）幾何学形状ベースのMVS概念を利用して、アルゴリズムを誘導し、解釈可能性を改良し、（d）疎／稠密技法の正確度および効率性から利益を享受するように、エンドツーエンド様式において、疎／稠密公式と併せて学習される。方法内で使用されるネットワークは、RGB画像のために1つと、疎深度画像のために1つとの、2つのエンコーダと、着目点検出のために1つと、記述子のために1つと、稠密深度予測のために1つとの、3つのデコーダとから成る、エンコーダーデコーダ構造で構成される、マルチタスクモデル（[参考文献22]参照）である。本方法また、効率的に、幾何学的初期値を使用して、点を三角測量し、着目点デコーダと、記述子デコーダと、疎深度エンコーダとの間の重要な連結を形成し、エンドツーエンド訓練を有効にする、可微分モジュールを利用する。

【0054】

場面の深度推定のための方法110およびシステム110の一実施形態は、広義には、図2の概略図に図示されるように、3つのステップに細分割されることができる。方法110は、広義には、図2に図示されるように、3つのステップに細分割されることができる。第1のステップ112では、標的またはアンカ画像114およびマルチビュー画像116が、共有RGBエンコーダおよび記述子デコーダ118（RGB画像エンコーダ119、検出器デコーダ121、および記述子デコーダ123を含む）を通して通過され、画像114、116毎に、記述子フィールド120を出力する。着目点122もまた、標的またはアンカ画像114に関して検出される。第2のステップ124では、アンカ画像114内の着目点122が、相対的姿勢126と併せて、代替視点から、基準画像116内の検索空間を決定するために使用される。記述子132は、それぞれ、エピソードサンプル127および点サンプル129を使用して、検索空間内でサンプリングされ、サンプリングされた記述子128を出力し、ソフトマッチャ130によって、着目点122に関する記述子128とマッチングされる。次いで、マッチングされた特徴点134は、三角測量モジュール136を使用する、SVDを使用して、三角測量され、3D点138を出力する。出力された3D点138は、疎深度エンコーダ140によって使用され、疎深度画

像を作成する。第3かつ最終のステップ142では、疎深度エンコーダ140に関する出力特徴マップおよびRGBエンコーダ119からの中間特徴マップが、集合的に、深度デコーダ144に通知し、稠密深度画像146を出力するために使用される。3つのステップはそれぞれ、下記により詳細に説明される。

【0055】

上記に説明されるように、共有RGBエンコーダおよび記述子デコーダ118は、2つのエンコーダ、すなわち、RGB画像エンコーダ119および疎深度画像エンコーダ140と、3つのデコーダ、すなわち、検出器デコーダ121（着目点検出器デコーダ121とも称される）、記述子デコーダ123、および稠密深度デコーダ144（稠密深度予測器デコーダ144とも称される）とから成る。一実施形態では、共有RGBエンコーダおよび記述子デコーダ118は、完全畳み込みニューラルネットワークアーキテクチャのSuperPoint様（[参考文献9]参照）公式を備えてもよく、これは、完全分解能画像に作用し、固定長記述子を伴う着目点検出を生産する。モデルは、単一共有エンコーダに、入力画像次元を処理および低減させる。RGBエンコーダ119からの特徴マップは、2つのタスク特有のデコーダ「ヘッド」の中にフィードされ、これは、着目点検出および着目点記述に関する加重を学習する。SuperPointにおける着目点検出および記述の本共同公式は、検出および記述タスクおよび深度推定の下流タスクに関する算出量を共有することを有効にする。しかしながら、SuperPointは、グレースケール画像上で訓練され、高フレームレートビデオストリーム上での持続的姿勢推定のための着目点検出および記述に焦点が当てられており、故に、比較的浅層エンコーダを有する。対照的に、本方法は、十分なベースライン、その結果、後続フレーム間のより長いインターバルを伴う、画像シーケンスに着目する。さらに、疎点分析のために好適なSuperPointの浅層バックボーンは、稠密深度推定の我々の下流タスクにとって、限定された容量を有する。故に、浅層バックボーンは、ResNet-50（[参考文献16]参照）エンコーダと置換され、これは、効率性および性能を平衡させる。着目点検出器デコーダ121の出力分解能は、SuperPointのものと同一である。点マッチングのために重要な微細および粗大レベル画像情報を融合させるために、方法110は、記述子デコーダ123のために、U-Net（[参考文献36]参照）様アーキテクチャを利用してもよい。記述子デコーダ123は、SuperPointと同様に、画像分解能の1/8において、N次元記述子テンソル120を出力する。本アーキテクチャは、図3に図示される。着目点検出器ネットワークは、オリジナルSuperPointネットワークの出力を希釈することによって訓練され、記述子は、下記に説明されるマッチング公式によって訓練される。

【0056】

前のステップは、アンカ画像に関する着目点と、全ての画像、すなわち、アンカ画像および基準画像の完全セットに関する記述子とを提供する。方法110の次のステップ124は、点マッチングおよび三角測量を含む。単純なアプローチは、アンカ画像114の記述子フィールド120からサンプリングされる着目点122の記述子を各基準画像116内のあらゆる可能性として考えられる位置にマッチングさせるためのものとなるであろう。しかしながら、これは、算出上法外となる。故に、方法110は、幾何学的制約を呼び出し、検索空間を制限し、効率性を改良する。マルチビュー幾何学形状からの概念を使用して、方法100は、基準画像内のエピポーラ線に沿ってのみ検索する（[図14]参照）。エピポーラ線は、関係 $\mathbf{x} \mathbf{F} \mathbf{x}^T = 0$ （式中、 $\mathbf{x}$ は、画像内の点の集合である）を使用する、基本行列 $\mathbf{F}$ を使用して決定される。マッチングされる点は、理想的シナリオでは、エピポーラ線上にあることが保証される。しかしながら、完璧な姿勢を取得することの実践的限界は、小固定オフセットを両側に伴う、エピポーラ線に沿った検索につながる。さらに、エピポーラ線は、深度値に関して、 $-\infty$ から ∞ まで伸展する。検索空間は、エピポーラ線から実行可能深度感知範囲内にあるように制約され、サンプリングレートは、図4に図示されるように、実装目的のために、同一出力形状を伴う記述子フィールドを取得するために、本制限された範囲内で変動される。双線形サンプリングが、記述子フィールド

10

20

30

40

50

120内の所望の点において記述子を取得するために使用される。各着目点122の記述子は、図12の方程式(1)に図示され、また、下記に再現されるように、画像視点毎に、その対応するエピポーラ線に沿って、記述子フィールド120で畳み込まれる。

【数1】

$$C_{j,k} = \tilde{D}_j * D_k^E, \forall x \in \mathcal{E}, \quad (1)$$

式中、 $\tilde{D}$ は、アンカ画像の記述子フィールドであり、 D^k は、節点jに関する固定されたエピポーラ線Eに沿って、全てのサンプリングされた点xにわたって畳み込まれる、k番目の基準画像の記述子フィールドである。これは、事実上、記述子フィールドと着目点記述子との間の相互相関マップ([参考文献2]参照)を提供する。本マップ内の高値は、アンカ画像内の着目点に対する基準画像内の潜在的特徴点マッチングを示す。実践では、我々は、訓練を容易にするために、バッチ正規化([参考文献20]参照)およびReLU非線形性([参考文献23]参照)を出力 $C_{j,k}$ に追加する。

10

【0057】

3D点を取得するために、[参考文献21]において提案される、代数三角測量アプローチが、追従される。各着目点jが、相互から独立して処理される。アプローチは、各相互相関マップ内のピーク値から取得される2D位置とともに、2D着目点を三角測量することに応じて、構築される。2D位置を推定するために、空間軸を横断して、softmaxが、図13の方程式(2)に図示され、また、下記に再現されるように、最初に算出される。

20

【数2】

$$C'_{j,k} = \exp(C_{j,k}) / \left(\sum_{r_x=1}^W \sum_{r_y=1}^H \exp(C_{j,k}(r_x, r_y)) \right), \quad (2)$$

30

式中、 $C_{j,k}$ は、j番目の中点およびk番目のビューに関する相互相関マップを示し、W、Hは、エピポーラ検索線の空間寸法である。次いで、節点の2D位置が、soft-argmax演算とも称される、対応する相互相関マップの質量中心として、計算される。

【0058】

次いで、図14の方程式(3)（また、下記に再現される）を使用して、節点の2D位置が、soft-argmax演算とも称される、対応する相互相関マップの質量中心として、計算される。

40

【数3】

50

$$x_{j,k} = \sum_{r_x=1}^W \sum_{r_y=1}^H r(x,y) (C'_{j,k}(r(x,y))) \quad (3)$$

soft-argmaxの重要な特徴は、最大値のインデックスを得るのではなく、勾配がマッチングされた点 $x_{j,k}$ の出力された2D位置から相互相関マップ $C_{j,k}$ にフーバックすることを可能にすることである。換言すると、argmaxと異なり、soft-argmax演算子は、可微分である。節点の3D位置をその2D推定値 $x_{j,k}$ から推測するために、線形代数三角測量アプローチが、使用される。本方法は、点 z_j の3D座標を見出すステップを、点 $\bar{z}$ のホモニアス3D座標ベクトルの方程式の優決定系を解くステップに還元する。

10

【0059】

soft-argmaxの重要な特徴は、最大値のインデックスを得るのではなく、勾配がマッチングされた点 $x_{j,k}$ の出力された2D位置から相互相関マップ $C_{j,k}$ にフーバックすることを可能にすることである。換言すると、argmaxと異なり、soft-argmax演算子は、可微分である。節点の3D位置をその2D推定値 $x_{j,k}$ から推測するために、線形代数三角測量アプローチが、使用される。本方法は、図15の方程式4に図示され、また、下記に再現されるように、点 z_j の3D座標を見出すステップを、点

20

【化1】

$\bar{z}$

のホモニアス3D座標ベクトルの方程式の優決定系を解くステップに還元する。

【数4】

$$A_j \bar{z} = 0, \quad (4)$$

式中、 $A_j \in \mathbb{R}^{2 \times 4}$ は、完全投影行列および $x_{j,k}$ からの成分から成る、行列である。単純な三角測量アルゴリズムは、各ビューからの点座標が、相互から独立し、したがって、全て、同等の寄与を三角測量に行うと仮定する。しかしながら、あるビューでは、2D点場所は、確実に推定されることができず(例えば、オクルージョン、モーションアーチファクト等に起因して)、最終三角測量結果の不必要な劣化につながる。これは、代数再投影誤差を最適化し、非均一注意を異なるビューに払う、方法の傾向を大幅に悪化させる。本問題は、RANSACを、Huber損失(正対応に対応する、再投影誤差をスコア化するために使用される)とともに適用することによって、解決されることができる。しかしながら、これは、その独自の短所を有する。例えば、RANSACを使用することは、除外されたビューへの勾配フローを完全に排除し得る。前述の問題に対処するために、加重 w_k が、異なるビューに対応する行列の係数に追加される。

30

40

【0060】

単純な三角測量アルゴリズムは、各ビューからの点座標が、相互から独立し、したがって、全て、同等の寄与を三角測量に行うと仮定する。しかしながら、あるビューでは、2D点場所は、確実に推定されることができず(例えば、オクルージョン、アーチファクト等に起因して)、最終三角測量結果の不必要な劣化につながる。これは、代数再投影誤差を最適化し、非均一注意を異なるビューに払う、方法の傾向を大幅に悪化させる。本問題は、ランダムサンプルコンセンサス(RANSAC)を、Huber損失(正対応に対応する、再投影誤差をスコア化するために使用される)とともに適用することによって、解

50

決されることができる。しかしながら、これは、その独自の短所を有する。例えば、RANSA Cを使用することは、除外されたビューへの勾配フローを完全に排除し得る。前述の問題に対処するために、加重 w_k が、図16の方程式(5)に図示されるように、異なるビューに対応する行列の係数に追加される。加重 w は、各相互相関マップ内の最大値であるように設定される。これは、着目点を三角測量する間、各カメラビューの寄与がより低く加重されることを可能にする。着目点の信頼度値は、1に設定されることに留意されたい。下記に再現される、図16の方程式(5)は、行列 $B = UDV^T$ の可微分特異値分解(SVD)を介して解かれ、そこから

【化2】

$\bar{z}$

10

が、 V の最後の列として設定される。

【数5】

$$(w_j A_j) z_j = 0, \quad (5)$$

加重 w は、各相互相関マップ内の最大値であるように設定される。これは、着目点を三角測量する間、各カメラビューの寄与がマッチングの品質によって制御され、低信頼度マッチングが着目点の三角測量の間により低く加重されることを可能にする。着目点の信頼度値は、1に設定されることに留意されたい。上記方程式は、行列 $B = UDV^T$ の可微分特異値分解を介して解か

20

れ、そこから $\bar{z}$ が、 V の最後の列として設定される。 z の最終非ホモジニアス値は、ホモジニアス3D座標ベクトル $\bar{z}$ をその第4座標 $z = \bar{z}/(\bar{z})_4$ によって除算することによって取得される。

【0061】

z の最終非ホモジニアス値は、ホモジニアス3D座標ベクトル

【化3】

$\bar{z}$

30

をその第4座標

【化4】

$z = \bar{z}/(\bar{z})_4$

によって除算することによって取得される。

【0062】

次に、疎深度点の緻密化を含む、方法110のステップ142が、説明されるであろう。特徴点検出器ネットワークが、点の位置を提供する。三角測量された点の z 座標は、深度を提供する。入力画像と同一分解能の疎深度画像が、これらの疎点の深度を用いて代入される。勾配は、疎深度画像から入力画像までの3D特徴点に逆伝搬し得ることに留意されたい。これは、SegNet([参考文献1]参照)におけるスイッチアンブリングに類似する。疎深度画像は、画像エンコーダネットワーク119のより限られたバージョンである、エンコーダネットワークを通して通過される。より具体的には、チャンネル幅を伴う、ResNet-50エンコーダが、画像エンコーダの1/4であるように、各層の後に使用される。これらの特徴は、画像エンコーダ119から取得される特徴と連結される。画像およびデコーダ内の同一分解能の中間特徴マップと連結された疎深度エンコーダの両方からの中間特徴マップを伴う、UNETスタイルデコーダが、[参考文献6]と同様に、使用される。4つのスケールにわたる深層監視が、提供される([参考文献25]参照)。空間ピラミッドプーリングブロックもまた、含まれ、異なる受容野サイズにおける特徴混合を促す([参考文献15、4]参照)。本アーキテクチャの詳細は、図5に

40

50

示される。

【0063】

全体的訓練目的が、ここで説明されるであろう。ネットワーク全体が、(a) 着目点検出器デコーダの出力テンソルと *Super Point* から取得されるグラウンドトゥールース着目点場所との間のクロスエントロピー損失と、(b) *soft-argmax* およびグラウンドトゥールース2D点マッチング後の2D点出力間の平滑-L1損失と、(c) *SV* 三角測量後の3D点出力とグラウンドトゥールース3D点との間の平滑-L1損失と、(d) 出力稠密深度マップ上の縁を意識した平滑度損失と、(e) 予測される稠密深度マップ出力とグラウンドトゥールース3D深度マップとの間の複数のスケールにわたる平滑-L1損失との組み合わせを用いて、訓練される。全体的訓練目的は、以下である。

10

【数6】

$$L = w_{ip}L_{ip} + w_{2d}L_{2d} + w_{3d}L_{3d} + w_{sm}L_{sm} + \sum_i w_{d,i}L_{d,i}, \quad (6)$$

L_{ip} は、着目点検出損失であり、 L_{2d} は、2Dマッチング損失であり、 L_{3d} は、3D三角測量損失であり、 L_{sm} は、平滑度損失であり、 $L_{d,i}$ は、オリジナル画像分解能から1/16の画像分解能にわたる4つの異なるスケールに関するスケールiにおける深度推定損失である。

20

【実施例】

【0064】

実施例：

【0065】

実装詳細：

【0066】

訓練：大部分のMVSデータセットは、DEMONデータセット上で訓練される。しかしながら、DEMONデータセットは、主に、関連付けられる深度および姿勢情報を伴う、対の画像を含有する。相対的信頼度推定は、我々のアルゴリズムにおける正確な三角測量に重要であって、信頼度を正確に推定し、着目点を全体的に三角測量するために、3つまたはそれを上回る長さのシーケンスを必要とする。故に、我々は、MVS深度推定のための従来のデータセットから退避し、代わりに、ScanNet（[参考文献8]参照）を使用する。ScanNetは、3Dカメラ姿勢、表面再構築物、およびインスタンスレベル意味論セグメント化を用いて注釈が付けられた、1,500を上回る走査内の250万ビューを含有する、RGB-Dビデオデータセットである。20フレームの固定インターバルにおける走査からの3つのビューが、姿勢および深度情報とともに、我々の方法における訓練データ点を形成する。標的フレームは、着目点を検出するために、*Super Point*を通して通過され、これは、次いで、我々のネットワークを訓練しながら、損失 L_{ip} を使用して希釈される。我々は、深度画像を使用して、グラウンドトゥールース2Dマッチングを決定し、深度を投影解除し、グラウンドトゥールース3D点を決定する。我々は、24のバッチサイズを伴う、PyTorchフレームワークと、学習率0.0001（ $\beta_1=0.9$ 、 $\beta_2=0.999$ ）を伴う、ADAMオプティマイザとを使用して、10万回の反復にわたって、我々のモデルを訓練する。我々は、画像の分解能をqVGA（ 240×320 ）および着目点の数を各画像内で512に固定し、着目点の最大で半分が、 $5e-4$ で閾値化された着目点検出器から選定され、点の残りは、画像からランダムに選定される。ランダム点を選定することは、画像内の疎点の均一分布を確実にし、高密度化プロセスに役立つ。我々は、マッチングが25と同程度に小さい長さに関してさえロバストであることが見出したにもかかわらず、エピソードラ線に沿ったサンプリングされる記述子の長さを100であるように設定する。我々は、実験的に、加重を[0.1, 1.0, 2.0, 1.0, 2.0]であるように設定する。

30

40

50

【0067】

評価：ScanNet試験セットは、訓練データセット内の707の場面に関して異なる一意の場面の100走査から成る。我々は、最初に、ScanNet上での姿勢推定の目的のために、我々の検出器および記述子デコーダの性能を評価する。我々は、SuperPointにおいて提案される評価プロトコルおよびメトリック、すなわち、平均位置特定誤差(MLE)、マッチングスコア(MScore)、再現性(Rep)、および回転に関して 5° (5度) 閾値および平行移動に関して5cmにおいて、記述子マッチングおよびPnPアルゴリズムを使用して推定される、正しい姿勢の割合を使用する。我々は、SuperPointの論文において提案されるように、Rep、MLE、およびMScoreに関して3ピクセルのNMS閾値において、SuperPoint、SIFT、ORB、およびSURFに対して比較する。次に、我々は、標準メトリックを使用して、我々の推定される深度の品質を定量的に測定する、すなわち、絶対相対的誤差(AbsRel)、絶対差誤差(Absdiff)、相対二乗誤差(SqRel)、二乗平均平方根誤差、およびその対数スケール(RMSEおよびRMSE対数)、および正対応比率($\delta < 1.25i$ 、式中、 $i \in 1, 2, 3$)である。

10

【0068】

我々は、我々の方法とMVSに関する最近の深層学習アプローチ、すなわち、(a) DPSTNet：深層平面掃引アプローチ、(b) MVDepthNet：マルチビュー深度ネットワーク、および(c) ガウスプロセスを使用する、GP-MVSTNet一時非パラメータ融合アプローチを比較する。これらの方法は、従来の幾何学形状ベースのステレオアルゴリズムよりはるかに良好に性能を発揮することに留意されたい。我々の一次結果は、長さ3のシーケンスに基づくが、また、場面長の関数としての性能を理解するために、長さ2、4、5、および7のシーケンスに基づく数も報告する。我々は、我々のアプローチの他の屋内場面への一般化を理解するために、Sun3Dデータセット上での方法の評価する。我々はまた、ランタイムにおける動作効率性を理解するために、異なる方法に関する乗算演算(MAC)についても議論する。

20

【0069】

記述子品質：

【0070】

図8における表1は、我々の検出器および記述子評価の結果を示す。MLEおよび再現性は、検出器メトリックであって、MScoreは、記述子メトリックであって、回転@5および平行移動@5は、組み合わせられたメトリックであることに留意されたい。我々は、我々の検出器に関する閾値を、訓練の間に使用されたものと同一である、0.0005に設定する。これは、多数の着目点が検出される結果をもたらす(Num)、これは、我々の発明に有利にあるように、再現性スコア(Rep)を人為的に上昇させるが、MLEメトリックによって示されるように、不良位置特定性能を有する。しかしながら、我々のMScoreは、完全画像ではなく、エッジ線に沿ってのみマッチングするように我々のネットワークを訓練したにもかかわらず、SuperPointと同等である。さらに、我々は、最良回転@5および平行移動@5メトリックを有し、我々の記述子を使用して見出されるマッチングが、回転および平行移動、すなわち、姿勢を正確に決定することに役立つことを示す。これらの結果は、我々の訓練プロシージャが、SuperPointのホモグラフィ適格技法を補完し、全体的性能を後押しし得ることを示す。

30

40

【0071】

深度結果：

【0072】

我々は、全てのシナリオに関して、かつ全てのデータセットを横断して、我々のネットワークを評価するために、同一ハイパーパラメータを設定する、すなわち、検出される点の数を512、サンプリングされる記述子の長さを100、および検出器閾値を $5e-4$ に固定する。着目点の均一分布を確実にし、クラスタを回避するために、我々は、[参考文献9]に提案されるように、9の高NMS値を設定する。付録は、ハイパーパラメータ

50

の異なる選択肢にわたるアブレーション研究を有する。図9の表2は、訓練セット内で使用されるように、長さ3および間隙20のシーケンス上での深度推定の性能を示す。正当な比較のために、我々は、競合アプローチの2つのバージョン、すなわち、(1)我々が提供したオープンソース訓練モデルと、(2)原稿内に提案された、または著者によって利用可能にされた、デフォルト訓練パラメータを用いて、10万回の反復にわたって、ScanNet上で微調整された、訓練モデルとを評価する。我々は、我々の発明と同様に、20フレームの間隙を使用して、各ネットワークを訓練する。微調整されたモデルは、図9の表2内の添字「FT」によって示される。当然のことながら、微調整されたモデルは、ScanNet評価上でのオリジナルモデルよりはるかに良好であった。MVDepthNetは、微調整後、少なくとも改良を有し、これは、訓練の間に使用される、著しい幾何学および測光拡張に起因し、故に、良好に一般化させることができる。DPSSNetは、絶対誤差における25%超の低下を伴って、微調整から最大限に利益を享受する。しかしながら、本開示の方法による、我々のネットワークは、全てのメトリックを横断して、全ての方法より優れている。図6は、異なる方法間の定質的比較を示し、図7は、推定される深度マップからの場面のサンプル3D再構築物を示す。

【0073】

任意のマルチビューステレオ方法の重要な特徴は、より多くのビューに伴って改良する能力である。図10の表3は、異なる数の画像に関する性能を示す。我々は、それぞれ、2、4、5、および7つのフレームに関して、フレーム間隙を20、15、12、および10に設定する。これらの間隙は、各集合が3D空間内の類似体積に近似的に及び、任意の性能改良が、新しい情報を入手することとは対照的に、利用可能な情報を使用して、ネットワークからより良好に出現することを確実にする。我々は、再び、本明細書に開示される方法が、異なるシーケンス長に関する全3つのメトリックにおいて、全ての他の方法より優れていることを認める。値のより詳細な検討は、DPSSNetおよびGPMVSSNetが、付加的ビューから利益を享受しない一方、MVDepthNetが、少数の付加的ビューから利益を享受するが、4つを上回るフレームに関して停滞することを示す。対照的に、本開示の方法は、付加的ビューを用いて、全3つのメトリックにおいて、定常改良を示す。これは、付加的ビューから必然的に利益を享受する、我々の点マッチャおよび三角測量モジュールに起因し得る。

【0074】

最終実験として、我々は、我々のネットワークを80対の画像から成るSun3D試験データセット上で試験する。Sun3Dはまた、ScanNetと比較してはるかに小さいスケールにあるにもかかわらず、屋内環境を捕捉する。図11の表4は、開示される実施形態による、我々のネットワークと比較した、前述で議論されるDPSSNetおよびMVDepthNetの2つのバージョンからの性能を示す。DPSSNetおよびMVDepthNetは、Sun3D訓練データベース上で元々訓練されていることに留意されたい。DPSSNetの微調整されたバージョンは、ScanNet訓練データベース内により広い多様性に起因して、Sun3D試験セット上でのオリジナルネットワークより良好に性能を発揮する。MVDepthNetは、対照的に、より不良に性能を発揮し、ScanNetに過剰適合し、オリジナルネットワークが、十分に訓練され、良好に一般化されたことを示す。着目すべきこととして、本明細書に開示される実施形態による、我々の方法は、再び、我々の訓練されたネットワークがSun3Dデータベースからの任意の画像を経ていないにもかかわらず、両方の方法より優れている。これは、疎深度を決定し、次いで、稠密化する、我々の原理方法が、良好な一般化可能性を有することを示す。

【0075】

次に、我々は、開示される実施形態による、我々のアプローチのために必要とされる、乗累算演算(MAC)の総数を評価する。2画像シーケンスに関して、我々は、点検出器および記述子モジュールに関して、16.57ギガマック(GMac s)、マッチャおよび三角測量モジュールに関して、0.002GMac s未満、および疎／稠密モジュールに関して、67.90GMac sを実施する。このうちの大部分は、画像および疎深度工

ンコーダをデコーダに接続する、U-NETスタイル特徴テンソルに起因する。我々は、合計84.48GMac sを実施し、2画像シーケンスに関する深度を推定する。これは、2画像シーケンスに関して295.63GMac sを実施する、DPSNetより著しく低く、また、深度を推定するために一对の画像に関して134.8GMac sを実施する、リアルタイムMVDepthNet未満である。深度をNvidia Titan RTX GPU上で推定するために、90ミリ秒かかり、これは、DPSNetより2.5倍高速であると評価される。我々は、我々の本開示の方法が、PyTorchの本来のSVDと三角測量のためのカスタム実装を置換することによって、さらに加速され得ると考える。さらに、コスト体積に依存しないため、画像認識およびオブジェクト検出のために導出されるものとしての複合スケールリング則は、我々の方法をより効率的にするために簡単に拡張されることができる。

10

【0076】

深度推定のための本開示の方法は、マルチビューステレオシナリオにおいて、疎点を三角測量および稠密化するように学習することによって、効率的深度推定アルゴリズムを提供する。既存のベンチマークの全てにおいて、本明細書に開示される方法は、最先端結果を上回っており、かつ競合的方法の有意な算出効率性を実証している。これらの方法は、着目点マッチングのためのより効果的注意機構およびさらなるアンカサポートビュー選択を組み込むことによって、拡張され得ることが予期される。本方法はまた、深度推定およびSLAMが相互の重複となるように、SLAM問題とのより深層の統合を組み込んでもよい。

20

【0077】

付属1：下記に列挙された参考文献は、上記の括弧内の参考文献（[参考文献##]）に対応し、これらの参考文献はそれぞれ、参照することによってその全体として本明細書に組み込まれる。

30

40

50

【表 1 - 1】

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation (2015)
2. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: Fully-convolutional siamese networks for object tracking. In: European conference on computer vision. pp. 850–865. Springer (2016)
3. Chang, J.R., Chen, Y.S.: Pyramid stereo matching network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5410–5418 (2018)
4. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
5. Chen, R., Han, S., Xu, J., Su, H.: Point-based multi-view stereo network. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1538–1547 (2019)
6. Chen, Z., Badrinarayanan, V., Drozdov, G., Rabinovich, A.: Estimating depth from rgb and sparse sensing. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 167–182 (2018)
7. Cheng, X., Wang, P., Yang, R.: Depth estimation via affinity learned with convolutional spatial propagation network. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 103–119 (2018)
8. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proc. Computer Vision and Pattern Recognition (CVPR), IEEE (2017)
9. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 337–33712 (June 2018). <https://doi.org/10.1109/CVPRW.2018.00060>
10. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE international conference on computer vision. pp. 2650–2658 (2015)
11. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2002–2011 (2018)
12. Garg, R., BG, V.K., Carneiro, G., Reid, I.: Unsupervised cnn for single view depth estimation: Geometry to the rescue. In: European Conference on Computer Vision. pp. 740–756. Springer (2016)
13. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 270–279 (2017)
14. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
15. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence **37**(9), 1904–1916 (2015)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
17. Hou, Y., Kannala, J., Solin, A.: Multi-view stereo by temporal nonparametric fusion. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2651–2660 (2019)

10

20

30

40

50

【表 1 - 2】

18. Huang, P.H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.B.: Deepmvs: Learning multi-view stereo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2821-2830 (2018)
19. Im, S., Joon, H.G., Lin, S., Kwon, I.S.: Dpsnet: End-to-end deep plane sweep stereo. In: 7th International Conference on Learning Representations, ICLR 2019. International Conference on Learning Representations, ICLR (2019)
20. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
21. Isakov, K., Burkov, E., Lempitsky, V., Malkov, Y.: Learnable triangulation of human pose. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 7718-7727 (2019)
22. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7482-7491 (2018)
23. 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)
24. Lasinger, K., Ranftl, R., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. arXiv preprint arXiv:1907.01341 (2019)
25. Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: Artificial intelligence and statistics. pp. 562-570 (2015)
26. Lee, J.H., Han, M.K., Ko, D.W., Suh, H.H.: From big to small: Multi-scale local planar guidance for monocular depth estimation. arXiv preprint arXiv:1907.10326 (2019)
27. Liu, Y., Shen, Z., Lin, Z., Peng, S., Bao, H., Zhou, X.: Gift: Learning transformation-invariant dense visual descriptors via group cns. In: Advances in Neural Information Processing Systems. pp. 6990-7001 (2019)
28. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91-110 (2004)
29. Ma, F., Cavallheiro, G.V., Karaman, S.: Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 3288-3295. IEEE (2019)
30. Ma, F., Karaman, S.: Sparse-to-dense: Depth prediction from sparse depth samples and a single image (2018)
31. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: Orb-slam: a versatile and accurate monocular slam system. IEEE transactions on robotics **31**(5), 1147-1163 (2015)
32. Murthy Jatavallabhula, K., Iyer, G., Paull, L.: gradslam: Dense slam meets automatic differentiation. arXiv preprint arXiv:1910.10672 (2019)
33. Nistér, D., Naroditsky, O., Bergen, J.: Visual odometry. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. vol. 1, pp. 1-1. Ieee (2004)
34. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652-660 (2017)
35. Riegler, G., Osman Ulusoy, A., Geiger, A.: Octnet: Learning deep 3d representations at high resolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3577-3586 (2017)

10

20

30

40

50

【表 1 - 3】

36. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234-241. Springer (2015)
37. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: 2011 International conference on computer vision. pp. 2564-2571. Ieee (2011)
38. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning feature matching with graph neural networks. arXiv preprint arXiv:1911.11763 (2019)
39. Sinha, A., Unimesh, A., Huang, Q., Ramani, K.: Surfnet: Generating 3d shape surfaces using deep residual networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6040-6049 (2017)
40. Wang, K., Shen, S.: Mvdepthnet: real-time multiview depth estimation neural network. In: 2018 International Conference on 3D Vision (3DV). pp. 248-257. IEEE (2018)
41. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 767-783 (2018)
42. Yi, K.M., Trulls, E., Lepetit, V., Fua, P.: Lift: Learned invariant feature transform. In: European Conference on Computer Vision. pp. 467-483. Springer (2016)
43. Zhang, Y., Funkhouser, T.: Deep depth completion of a single rgb-d image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 175-185 (2018)
44. Zhang, Z.: Microsoft kinect sensor and its effect. IEEE multimedia 19(2), 4-10 (2012)

10

20

【0078】

本発明の種々の例示的实施形態が、本明細書に説明される。これらの実施例は、非限定的意味で参照される。それらは、本発明のより広く適用可能な側面を例証するために提供される。種々の変更が、説明される本発明に行われてもよく、本発明の真の精神および範囲から逸脱することなく、均等物が代用されてもよい。加えて、多くの修正が、特定の状況、材料、組成物、プロセス、プロセス作用、またはステップを本発明の目的、精神、または範囲に適合させるために行われてもよい。さらに、当業者によって理解されるであろうように、本明細書で説明および例証される個々の変形例はそれぞれ、本発明の範囲または精神から逸脱することなく、他のいくつかの実施形態のうちのいずれかの特徴から容易に分離される、またはそれらと組み合わせられる、離散コンポーネントおよび特徴を有する。全てのそのような修正は、本発明と関連付けられる請求項に記載の範囲内であることが意図される。

30

【0079】

本発明は、本主題のデバイスを使用して実施され得る方法を含む。本方法は、そのような好適なデバイスを提供する行為を備えてもよい。そのような提供は、エンドユーザによって実施されてもよい。換言すると、「提供する」行為は、単に、エンドユーザが、本主題の方法において必要なデバイスを取得する、それにアクセスする、それに接近する、それを位置付ける、それを設定する、それを活性化する、それに電源を入れる、または別様にそれを提供するように作用することを要求する。本明細書に列挙される方法は、論理的に可能な列挙されたイベントの任意の順序およびイベントの列挙された順序で行われてもよい。

40

【0080】

本発明の例示的側面が、材料選択および製造に関する詳細とともに、上記に記載されている。本発明の他の詳細に関して、これらは、前述の参照特許および刊行物に関連して理解され、概して、当業者によって公知である、または理解され得る。同じことが、一般的または理論的に採用されるような付加的行為の観点から、本発明の方法ベースの側面に関

50

して当てはまり得る。

【0081】

加えて、本発明は、随意に、種々の特徴を組み込む、いくつかの実施例を参照して説明されているが、本発明は、発明の各変形例に関して検討されるように説明または図示されるものに限定されるものではない。種々の変更が、説明される本発明に行われてもよく、均等物（本明細書に列挙されるか、またはある程度の簡潔目的のために含まれないかどうかにかかわらず）は、本発明の精神および範囲から逸脱することなく代用されてもよい。加えて、値の範囲が提供される場合、その範囲の上限と下限との間の全ての介在値および任意の他の述べられた値または述べられた範囲内の介在値が、本発明内に包含されるものと理解されたい。

10

【0082】

また、説明される本発明の変形例の任意の随意の特徴は、独立して、または本明細書に説明される特徴のうちの任意の1つまたはそれを上回るものと組み合わせて、記載および請求され得ることが検討される。単数形の項目の言及は、複数の同一項目が存在する可能性を含む。より具体的には、本明細書および本明細書に関連付けられる請求項で使用されるように、単数形「a」、「an」、「said」、および「the」は、別様に具体的に述べられない限り、複数の言及を含む。換言すると、冠詞の使用は、上記の説明および本発明と関連付けられる請求項における本主題のアイテムのうちの「少なくとも1つ」を可能にする。さらに、そのような請求項は、任意の随意の要素を除外するように起草され得ることに留意されたい。したがって、本文言は、請求項の要素の列挙と関連する「単に」、「のみ」、および同等物等の排他的専門用語の使用、または「消極的」限定の使用のための先行詞としての役割を果たすことが意図される。

20

【0083】

そのような排他的専門用語を使用しなければ、本発明と関連付けられる請求項における用語「～を備える（comprising）」は、所与の数の要素が請求項で列挙されるかどうかにかかわらず、任意の付加的要素の包含を可能にするものとする、または特徴の追加は、そのような請求項に記載される要素の性質を変換すると見なされ得る。本明細書で具体的に定義される場合を除いて、本明細書で使用される全ての技術および科学用語は、請求項の正当性を維持しながら、可能な限り広い一般的に理解されている意味を与えられるべきである。

30

【0084】

本発明の範疇は、提供される実施例および／または本明細書に限定されるべきではなく、むしろ、本開示と関連付けられる請求項の用語の範囲のみによって限定されるものとする。

40

50

【図面】
【図 1】

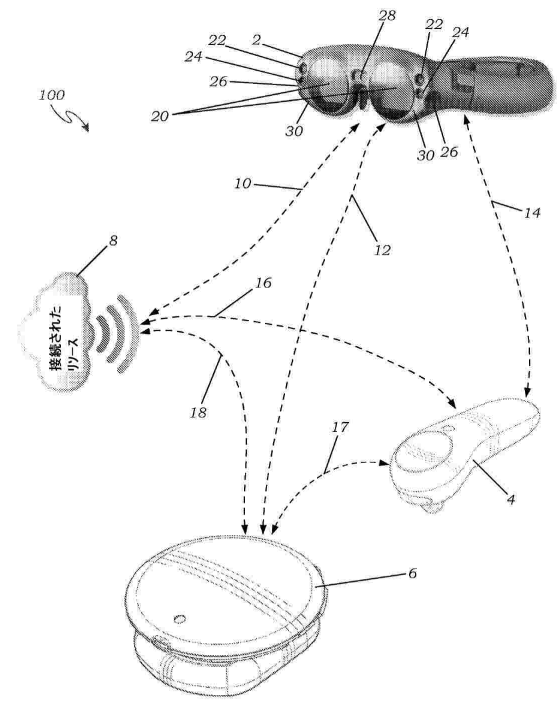


FIG. 1

【図 2】

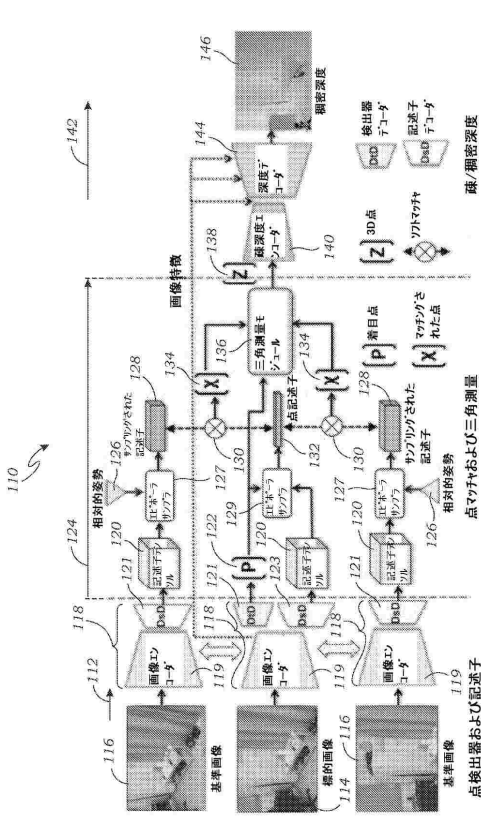
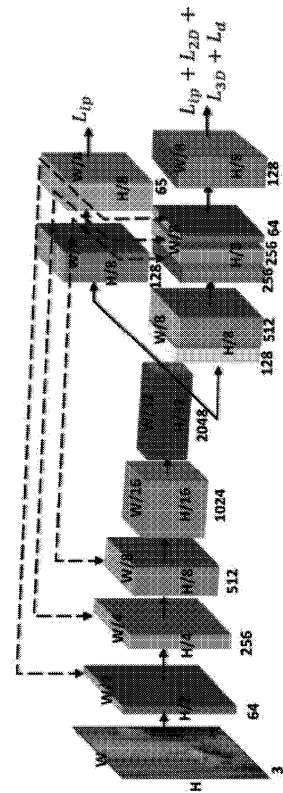


FIG. 2

【図 3】



【図 7】

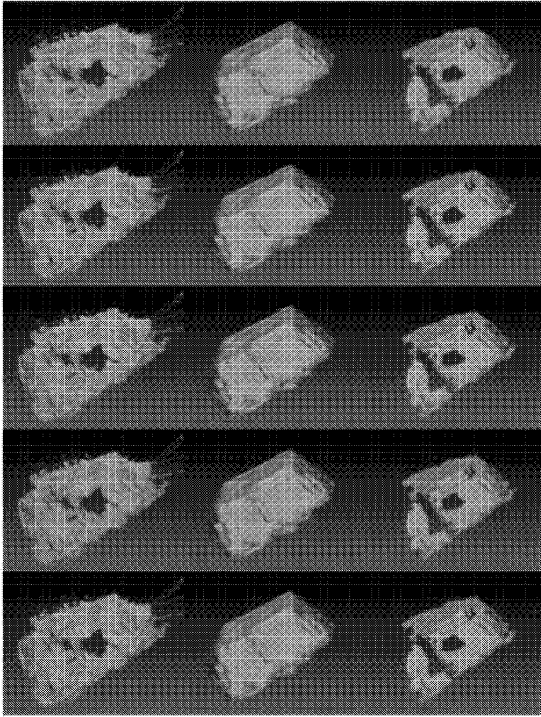


FIG. 7

【図 8】

	MLE	MScore	Num	Rep	rot@5	trans@5
ORB	2.584	0.194	401	0.613	0.142	0.064
SIFT	2.327	0.201	203	0.496	0.311	0.148
SURF	2.577	0.198	268	0.460	0.303	0.134
SP	2.545	0.375	129	0.519	0.489	0.244
Ours	3.101	0.329	1511	0.738	0.518	0.254

表 1 ScanNet 上での異なる記述子の性能

FIG. 8

【図 9】

	Abs-Rel	Abs	Sq-Rel	δ	δ^2	δ^3
GPMvsNet	0.1306	0.2600	0.0944	0.0841	0.9462	0.9753
GPMvsNet-FT	0.1079	0.2255	0.0960	0.8905	0.9591	0.9789
MVDepthNet	0.1191	0.2096	0.0910	0.8690	0.9599	0.9851
MVDepthNet-FT	0.1054	0.1911	0.0970	0.8952	0.9707	0.9895
DPSNet	0.1470	0.2248	0.1035	0.8486	0.9474	0.9761
DPSNet-FT	0.1025	0.1675	0.0574	0.9102	0.9708	0.9872
Ours	0.0932	0.1540	0.0506	0.9287	0.9767	0.9893

表2.ScanNet上での深度推定の性能(我々は、長さ3のシーケンスを使用して、207フレーム毎にサンプリングする)

FIG. 9

【図 10】

	2 フレーム			4 フレーム			5 フレーム			7 フレーム		
	AbR	Abs	SqR	AbR	Abs	SqR	AbR	Abs	SqR	AbR	Abs	SqR
GPN	0.112	0.233	0.101	0.109	0.226	0.100	0.107	0.226	0.112	0.109	0.230	0.116
MVN	0.126	0.238	0.471	0.105	0.191	0.078	0.106	0.192	0.071	0.108	0.195	0.067
DPS	0.099	0.181	0.062	0.102	0.168	0.057	0.102	0.168	0.057	0.102	0.167	0.057
Ours	0.106	0.173	0.057	0.090	0.150	0.049	0.088	0.147	0.048	0.087	0.144	0.043

表3.ScanNet上での深度推定の性能(我々は、異なる長さのシーケンスを使用する)

FIG. 10

10

20

30

40

50

【図 1 1】

	Abs-Rel	Abs	Sq-Rel	δ	δ^2	δ^3
MVDepthNet	0.1377	0.3199	0.1564	0.8245	0.9601	0.9851
MVDepthNet-FT	0.3092	0.7209	4.4899	0.7873	0.9117	0.9387
DPSNet	0.1590	0.3341	0.1564	0.8087	0.9363	0.9787
DPSNet-FT	0.1274	0.2858	0.0855	0.8396	0.9459	0.9866
Ours	0.1245	0.2662	0.0741	0.8551	0.9728	0.9902

表4. Sun3D上での深度推定の性能(我々は、長さ20のシーケンスを使用する)

FIG. 11

【図 1 2】

$$C_{j,k} = \hat{D}_j * D_j^k, \forall x \in \mathcal{E}, \tag{1}$$

式中、 $\hat{D}$ は、アノタ画像の記述子フィールドであって、 D^k は、節点*j*に関する固定されたエッジ-ラベルに沿って、全てのサンプリングされた点*x*にわたって畳み込まれる、*k*番目の基準画像の記述子フィールドである。これは、事実上、記述子フィールドと着目点記述子との間の相互相関マップ([参考文献2]参照)を提供する。本マップ内の高値は、アノタ画像内の着目点に対する基準画像内の潜在的特徴点マッチングを示す。実践では、我々は、訓練を容易にするために、バッチ正規化([参考文献20]参照)およびReLU非線形性([参考文献23]参照)を出力*C_{j,k}* に追加する。

FIG. 12

10

20

【図 1 3】

$$C'_{j,k} = \exp(C_{j,k}) / (\sum_{r_x=1}^W \sum_{r_y=1}^H \exp(C_{j,k}(r_x, r_y))), \tag{2}$$

式中、*C_{j,k}*は、*j*番目の中点および*k*番目のエッジに関する相互相関マップを示し、*W*, *H*は、エッジ-ラベル検索線の空間寸法である。

FIG. 13

【図 1 4】

$$x_{j,k} = \sum_{r_x=1}^W \sum_{r_y=1}^H r(x,y) (C'_{j,k}(r(x,y))) \tag{3}$$

FIG. 14

30

40

50

【図 15】

$$A_j \bar{z}_j = 0,$$

(4)

式中、 $A_j \in R^{2k,4}$ は、完全投影行列および $x_{k+1,4}$ からの成分から成る、行列である。

FIG. 15

【図 16】

$$(w_j A_j) \bar{z}_j = 0,$$

(5)

式中、 w_k は、異なる $\bar{z}_j$ に対応する、行列の係数に対する加重である。加重 w は、各相互相関マップ内の最大値であるように設定される。

FIG. 16

10

20

30

40

50

フロントページの続き

- (74)代理人 100181641
弁理士 石川 大輔
- (74)代理人 230113332
弁護士 山本 健策
- (72)発明者 シンハ, アヤン トゥヒネンドゥ
アメリカ合衆国 カリフォルニア 9 4 1 5 8, サンフランシスコ, キング ストリート 3 0 1
- 審査官 高野 美帆子
- (56)参考文献 arxiv、Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, Long Quan, MVSNet: Depth Inference for Unstructured Multi-view Stereo, 外国雑誌, 発行日 20180717, US, <https://arxiv.org/pdf/1804.02505>
Daniel DeTone; Tomasz Malisiewicz; Andrew Rabinovich, SuperPoint: Self-Supervised Interest Point Detection and Description, 外国雑誌、2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), US, 発行日 20180618, 337-349, <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8575521>
- (58)調査した分野 (Int.Cl., D B名)
G 0 6 T 7 / 5 5
G 0 6 T 7 / 0 0
G 0 1 B 1 1 / 2 4 5