Abst

本研究では、1枚のRGBD画像から6DoF物体の姿勢を推定するデータ駆動型の手法を開発しました。従来の手法では、ポーズパラメータを直接回帰していましたが、本手法では、キーポイントベースのアプローチにより、この困難な課題に取り組みます。具体的には、物体の3Dキーポイントを検出し、6Dポーズパラメータを最小二乗法で推定するディープハフボーティングネットワークを提案する。 我々の手法は、RGBベースの6DoF推定に成功した2Dキーポイントアプローチの自然な拡張である。本手法は、RGBベースの6DoF推定に成功している2Dキーポイントアプローチを自然に拡張したものであり、奥行き情報を追加することで、剛体の幾何学的制約を完全に利用することができ、ネットワークの学習と最適化が容易である。実験の結果、我々の手法は、いくつかのベンチマークにおいて、最先端の手法を大幅に上回ることが示されました。コードとビデオは以下のサイトでご覧いただけます。

1. 緒言

本論文では，6DoFポーズ推定の問題を研究する．すなわち，正準フレームにおける物体の3次元位置と姿勢を認識することである．これは、ロボットによる把持・操作[6,48,55]、自律走行[11,5,53]、拡張現実[31]など、多くの実世界のアプリケーションにおいて重要な要素である。

6DoFの推定は、照明の変化、センサーノイズ、シーンのオクルージョン、オブジェクトの切り詰めなどにより、非常に困難な問題であることがわかっています。従来の手法[19,30]では，画像と物体のメッシュモデルとの対応関係を抽出するために，人間が作成した特徴量を用いていました．しかし、このような経験的に人間が作成した特徴量は、照明条件の変化やオクルージョンの多いシーンでは性能が低下してしまう。最近では，機械学習や深層学習技術の爆発的な発展に伴い，ディープニューラルネットワーク（DNN）に基づく手法がこのタスクに導入され，有望な改善が見られるようになった．[50,52]は、DNNで直接オブジェクトの回転と変換を回帰させることを提案した。しかし、これらの方法は、[37]で説明された回転空間の非線形性のために、一般化が不十分であった。その代わり，最近の研究では，DNNを利用して物体の2Dキーポイントを検出し，Perspective-n-Point（PnP）アルゴリズムを用いて6Dポーズパラメータを計算している[37,36,41,47]．これらの2段階のアプローチは、より安定した性能を発揮するが、そのほとんどが2Dプロジェクションの上に構築されている。**投影時には小さな誤差でも、実際の3D空間では大きな誤差となる可能性があります。また、2D投影後に3D空間の異なるキーポイントが重なってしまうことがあり、それらを区別することは困難である。さらに、剛体の幾何学的拘束情報は、投影によって部分的に失われてしまう。**

一方で，安価なRGBDセンサの開発により，より多くのRGBDデータセットが利用可能になっている．PointFusion[53]、Frustum pointnets[39]、VoteNet[38]のように、余分な奥行き情報があることで、2Dアルゴリズムを3D空間に拡張し、より良い性能を得ることができる。この目的のために，我々は2Dキーポイントベースのアプローチを3Dキーポイントに拡張し，剛体の幾何学的制約情報を完全に利用することで，6DoF推定の精度を大幅に向上させた．具体的には，図1に示すように，ポイント単位の3Dオフセットを学習し，3Dキーポイントに投票するディープ3Dキーポイントハフ投票ニューラルネットワークを開発しました． **3Dキーポイントとは，3次元空間における剛体の2点間の位置関係が固定されているという単純な幾何学的性質です。**したがって、物体表面上の可視点が与えられた場合、その座標と向きは深度画像から得られ、選択されたキーポイントへの並進オフセットも固定されており、学習可能です。一方、ポイント単位のユークリッドオフセットの学習は、ネットワークにとって簡単で、最適化も容易です。

また、複数のオブジェクトが存在するシーンを扱うために、インスタンスセマンティックセグメンテーションモジュールをネットワークに導入し、キーポイント投票と共同で最適化を行いました。その結果、これら

のタスクを共同で学習することで、お互いのパフォーマンスが向上することがわかりました。具体的には、意味的な情報は、ある点がどの部分に属するかを識別することで、トランスレーション・オフセットの学習を向上させ、トランスレーション・オフセットに含まれるサイズ情報は、外観は似ているがサイズが異なるオブジェクトを区別するのに役立ちます。

さらに，我々の手法を評価するために，YCB-VideoデータセットとLineMODデータセットを用いた実験を行った．実験の結果，我々の手法は，現在の最先端の手法よりもかなりのマージンで優れていることがわかった．

要約すると、この作品の主な貢献は以下の通りです。

- 単一RGBD画像の6DoFポーズ推定のための、インスタンスセマンティックセグメンテーションを備えた新しいディープ3Dキーポイントハフ投票ネットワーク。
- YCB および LineMOD データセットにおける最新の 6DoF ポーズ推定性能.
- 3D-keypointを用いた手法を詳細に分析し、従来の手法と比較した結果、6DoFポーズ推定の性能を向上させるためには、3D-keypointが重要な要素であることを示した。また、3D-keypointとセマンティックセグメンテーションを共同で学習することで、さらに性能が向上することを示しています。

2. 関連研究

2.1. ホリスティック・メソッド

ホリスティックな手法は、画像内の物体の3次元的な位置と向きを直接推定するものです。古典的なテンプレートベースの手法では，剛体のテンプレートを構築し，画像をスキャンして，最もマッチしたポーズを計算する[21,13,17]．このようなテンプレートは、クラスタ化されたシーンに対してロバストではない。最近、カメラや物体の6Dポーズを直接回帰するDNN（Deep Neural Network）ベースの方法がいくつか提案されている[52,50,14]。しかし、回転空間の非線形性により、データ駆動型DNNの学習と一般化は困難である。この問題に対処するために、いくつかのアプローチでは、ポーズを反復的に改良するためにpost-refinement手順[26,50]を使用し、他のアプローチでは、回転空間を離散させ、分類問題に単純化する[49,43,45]。後者のアプローチでは、離散化によって犠牲になった精度を補うために、ポスト・レフィンメント処理が依然として必要である。

2.2. キーポイントを使った手法

現在のキーポイントベースの手法は、まず画像内の物体の2Dキーポイントを検出し、次にPnPアルゴリズムを利用して6Dポーズを推定する。古典的な手法[30,42,2]は、豊富なテクスチャを持つオブジェクトの2Dキーポイントを効率的に検出することができる。しかし、これらの手法は、テクスチャのないオブジェクトを扱うことができません。深層学習技術の発展に伴い、ニューラルネットワークベースの2Dキーポイント検出法がいくつか提案されている。[41,47,20]では，キーポイントの2次元座標を直接回帰し，[33,24,34]では，ヒートマップを用いて2次元キーポイントを検出しています．また，[37]では，切り捨てられたシーンやオクルージョンされたシーンをうまく扱うために，2Dキーポイントの位置を投票するためのピクセル単位の投票ネットワークを提案している．これらの2Dキーポイントベースの手法は、オブジェクトの2D投影誤差を最小化することを目的としている。しかし，投影時には小さくても，実際の3D世界では大きな誤差が生じる可能性がある．[46]は，3Dポーズを復元するために，合成RGB画像の2つのビューから3Dキーポイントを抽出しています．しかし，これらはRGB画像しか利用していないため，剛体の幾何学的拘束情報が投影により部分的に失われ，また，3次元空間内の異なるキーポイントは，2次元に投影された後には重なってしまい，識別が困難

になる可能性がある．しかし、安価なRGBDセンサーの登場により、撮影した奥行き画像を使って3Dであらゆることができるようになりました。

2.3. 密な対応方法

これらのアプローチでは、Hough voting scheme [28,44,12]を利用して、ピクセルごとの予測で最終結果を投票する。これらの手法では、ランダムフォレスト[3,32]またはCNN[23,9,27,35,51]を用いて特徴を抽出し、各ピクセルに対応する3Dオブジェクト座標を予測し、最終的なポーズ結果を投票で決定する。このような高密度の2D-3D対応により、これらの手法は、出力空間が非常に大きくなるものの、オクルージョンのあるシーンに対してロバストになります。PVNet[37]では、2Dキーポイントに対してピクセル単位の投票を行い、Dense法とキーポイントベースの手法の利点を組み合わせています。さらに、この手法を追加の深度情報を持つ3Dキーポイントに拡張し、剛体の幾何学的制約を完全に利用します。

3.

6DoFポーズ推定の課題は，RGBD画像が与えられたときに，物体をその物体世界座標系からカメラ世界座標系に変換する剛体変換を推定することである。この変換は、3次元回転 $R \in SO(3)$ と並進 $t \in \boldsymbol{R}^3$ で構成される。

5. 結言

我々は、6DoFポーズ推定のためのインスタンスセマンティックセグメンテーションを用いた新しい深層3Dキーポイント投票ネットワークを提案し、いくつかのデータセットにおいて従来のアプローチよりも大きなマージンで優れた結果を得た。また、3Dキーポイントとセマンティックセグメンテーションを同時に学習することで、お互いの性能を高めることができることを示しています。3Dキーポイントに基づいたアプローチは、6DoFポーズ推定問題を解決するための有望な方向性であると考えています。