Aligning 3D Models to RGB-D Images of Cluttered Scenes

Abst

本研究の目的は、RGB-Dのシーンにあるオブジェクトを、ライブラリにある対応する3Dモデルで表現することです。本研究では、まずシーン内の物体を検出してセグメント化し、次に畳み込みニューラルネットワーク（CNN）を用いて物体の姿勢を予測することで、この問題に取り組んでいます。このCNNは、合成物体のレンダリングを含む画像のピクセル表面法線を用いて学習されます。 この方法は、実データでテストしたところ、実データで学習した他のアルゴリズムよりも優れていました。次に、この粗いポーズ推定値と推測されるピクセルサポートを用いて、少数のプロ・プロトタイプモデルをデータに合わせ、最もフィットしたモデルをシーンに配置します。 この結果、現在の最新技術[34]と比較して、3D検出タスクの性能が48%相対的に向上し、しかも1桁速いことが確認されました。

Intro

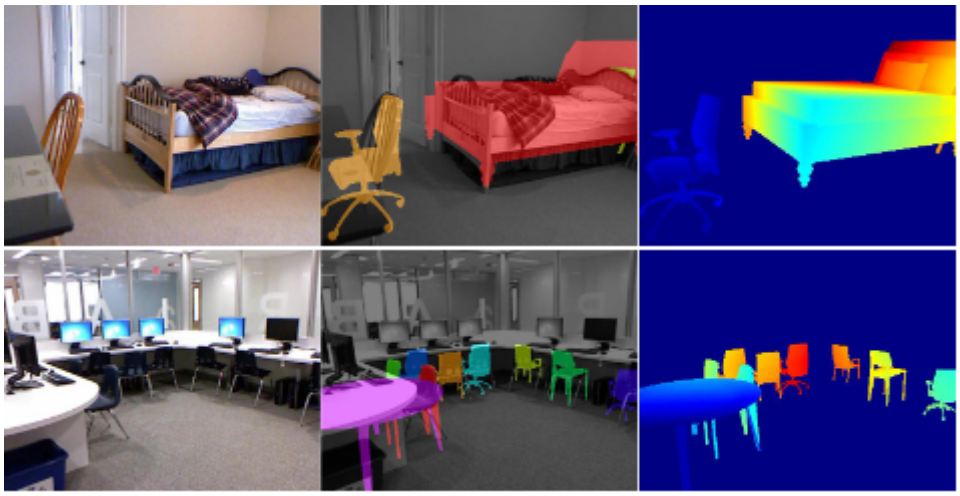

図1 本システムの出力。RGB画像をもとに、各オブジェクトが3Dモデルに置き換えられた3Dシーンを生成します。

シーンを真に理解するには、見えるものだけでなく、見えないものも含めて推論する必要があります。例えば、図1の画像を考えてみましょう。 椅子という物体を認識した後、その物体がどこまで奥行きがあるのか、別の視点からはどのように見えるのか、といったことをかなり理解しています。 こ

のような理解をコンピュータビジョンシステムで実現するためには、椅子の3DCADモデルをレンダリングすることで、椅子のピクセルを「再配置」することが考えられます。 このように3DCADモデルとの対応を明示的に行うことで、オブジェクト検出、セマンティックセグメンテーション、インスタンスセグメンテーション、きめ細かなカテゴリー化、ポーズ推定といった従来のコンピュータビジョンのアルゴリズムでは得られなかった豊かな表現が可能になります。 これらのタスクは、軌道最適化、動作計画、把持推定など、ロボット工学の観点からは十分ではありません。我々の提案するシステムは、散らかった屋内シーンの単一のRGB-D画像から始まり、図1に示すような出力を生成します。我々のアプローチは、以下のことを可能にします。

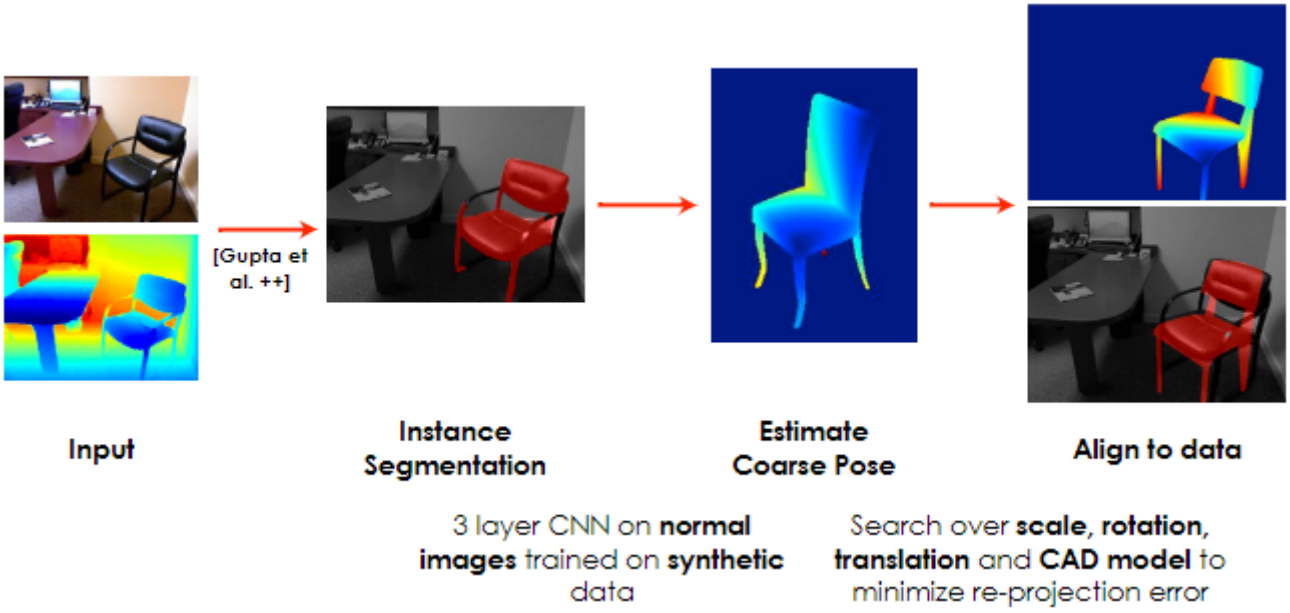

図2 アプローチの概要。 まず，[13]の物体検出インスタンス分離の出力を改良したものから始めます． まず，畳み込みニューラルネットワークを用いて物体の姿勢を推論し，データを説明する最適なモデルを探索する．

図2は我々のアプローチを示したものである． このCNNを、深度画像ではなく表面の法線画像を入力とする合成データで学習させます。その結果、合成データで学習したCNNは、実データで学習したCNNよりも優れた動作をすることがわかりました。 次に、推測されたポーズの仮説を用いて、3Dモデルの縮尺と正確な位置を含む小さなセットを検索する。 このタスクには、修正された it-erative closest point (ICP) アルゴリズムを使用し、適切に初期化された場合には、正確なインスタンスではなくオブジェクト・カテゴリのレベルで作業しても、妥当な結果が得られることを示す。これにより、すべてのモデルの学習には画像上の2Dアノテーションのみを使用し、テスト時にはシーンの豊富な3D表現を生成することができました。

最終的な出力は、画像内のオブジェクトに位置合わせされた3Dモデルのセットです。 我々のシステムからの出力の豊富さと質の高さは、現在の最新の3D検出方法と比較すると明らかになります。
我々の出力の自然な副産物として、シーン内の各オブジェクトの3Dバウンディングボックスがあります。この3Dバウンディングボックスを3D検出に使用した場合、現在の最新手法

（「SlidingShapes」）[34]と比較して、APポイントの絶対値で19%（相対値で48%）の改善が見られ、同時に少なくとも1桁以上の高速化が実現している。

3. Estimating Coarse Pose

本節では，深さ方向の画像から剛体の粗い姿勢を推定するための畳み込みニューラルネットワークを提案する．最近の研究[38]では，RGB画像の問題を研究している．

$C(k, n, s)$ はカーネルサイズk×k,n個のフィルターとs個のストライドを持つ畳み込み層、P{max,ave}(k,s)はカーネルサイズk×kとストライドを持つ最大値または平均値のプーリング層、Naは局所応答正規化層、RLAは整流線形ユニット、D(r)はドロップアウト率を持つドロップアウト層であると仮定すると、我々のネットワークは以下のアーキテクチャを持つ。