Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images

Abst

我々は，RGB-D画像におけるアモーダルな3Dオブジェクト検出のタスクに焦点を当て，オブジェクトの3Dバウンディングボックスをメトリック形式で最大限に生成することを目的としています。我々は，RGB-D画像からの3Dボリュームシーンを入力とし，3Dオブジェクトのバウンディングボックスを出力する3D ConvNetの定式化であるDeep Sliding Shapesを紹介する。我々のアプローチでは，幾何学的形状からオブジェクト性を学習するための初の3D領域提案ネットワーク（RPN）と，3Dの幾何学的特徴と2Dの色の特徴を抽出するための初の共同オブジェクト認識ネットワーク（ORN）を提案しています。特に，2つの異なるスケールでアモーダルRPNを学習し，3Dバウンディングボックスを回帰するORNを学習することで，様々なサイズのオブジェクトを扱う。実験によると，我々のアルゴリズムは，最先端のアルゴリズムよりもmAPで13.8倍，オリジナルのSliding Shapesよりも200倍高速であることがわかった。

Intro

一般的な物体検出では，物体のカテゴリと，その物体の可視部分の画像平面上の2次元バウンディングボックスを予測します。このような結果は，物体の検索などの一部のタスクには有効ですが，現実の3D世界に根ざした更なる推論を行うには，かなり物足りないものです。この論文では，RGB-D画像におけるアモーダル3Dオブジェクト検出のタスクに焦点を当てています。これは，オブジェクトの3Dバウンディングボックスを生成することを目的としており，切り捨てやオクルージョンに関わらず，オブジェクトの全範囲で実世界の寸法を与えることができます。このような認識は，例えばロボットのような知覚と操作のループにおいては，より有用です。しかし，予測のために新たな次元を追加すると，探索空間が大幅に拡大してしまい，作業の難易度が高くなります。

信頼性が高く手頃な価格のRGB-Dセンサー（Microsoft Kinectなど）が登場したことで，この重要なタスクを再検討する機会が訪れました。しかし，2次元の検出結果を単純に3次元に変換しても，うまくいきません（表3および[10]を参照）．そこで，奥行き情報を有効に利用するために，3D検出ウィンドウを3D空間内でスライドさせるSliding Shapes [25]が提案された．手作業で作成した特徴量を使用することには限界があるが，このアプローチはタスクを3Dで自然に定式化する。

一方，Depth RCNN [10]は，2Dのアプローチを採用しています。つまり，奥行きをカラー画像のエキストラチャンネルとして扱うことで，2D画像平面内のオブジェクトを検出し，ICPアライメントを使用して，検出された2Dウィンドウ内のポイントに3Dモデルをフィットさせます。現在，2D中心のDepth RCNNは，3D中心のSlid-ing Shapesを上回っています。 しかし，Depth RCNNの強みは，2D表現ではなく，ImageNetで事前に学習された優れた設計のディープネットワークを使用していることにあるのかもしれません。3Dでの深層学習を活用することで，エレガントでより強力な3D表現を得ることはできないでしょうか。