1P-05

Deep Learningを用いた 深度画像からの3Dコンテンツの自動生成

楊　夢龍†　　長尾　確†

名古屋大学 大学院情報学研究科†

1. はじめに

VRとAR技術の発展に伴い、2次元の写真やビデオではなく、3Dコンテンツの需要が爆発的に増加している。また、センサーと端末の小型化により、3Dスキャニング技術を利用し、人々は身近な物をモデリングし、VRやARに利用することが可能になった。しかし、既存のスキャニング手法ではオブジェクトを全方位からスキャニングする必要があり、多くの時間やコストがかかってしまう。

そこで本研究では、ディープラーニングの手法の一つである敵対的生成ネットワークを用いた、任意の角度から撮った一枚の深度画像に基づく3Dコンテンツ自動生成の手法を提案する。

2. 敵対的生成ネットワーク

敵対的生成ネットワークとは、2014年にIan Goodfellow[1]によって考察された、生成系ニューラルネットワークモデルの一種であり、目標を生成する生成モデルと、生成された物かどうかを判断する識別モデルがあり、両方が交替で訓練され、お互いに敵対することで生成モデルの精度を高めていく仕組みである。従来のAutoencoderのような生成モデルに対して、敵対的生成ネットワークの生成器は識別器から逆伝播した勾配を学習することで、より高いクオリティの生成が可能である。

3. データ収集と処理

本研究では、一枚の深度画像から、オブジェクトの背景もなく完全な3Dコンテンツを生成することを目標している。そのために、実世界におけるオブジェクトの深度画像を集める必要があり、集まったデータの前処理が必要である。本研究では椅子を対象として、30種類の椅子の59193枚の深度画像を収集した。

3.1. 深度画像の収集と3Dモデリング

深度画像の収集において、本研究ではマイクロソフト社のRGB-Dカメラ Kinect V2を利用し、様々な背景において、椅子を中心に周囲を回ってRGB-D画像を取得した。収集中にカメラとオブジェクトとの距離、角度を調整することで、1種類の椅子に対して多様なデータを収集できる。

本研究では3Dモデルを学習の真値(ground truth)とするので、より正確なオブジェクトモデルを作成することが必要である。そこで我々はOccipital社のStructure Sensorを使用し、Kinectのデータ収集とは別にオブジェクトのモデリングを行った。

3.2. データの前処理

本研究では、ニューラルネットワークの入力を2.5Dボクセルとし、出力の真値を3Dモデルのボクセルとした。そのために、収集したデータをボクセル化する必要がある。その流れを図1に示す。我々はPointCloud Libraryを用いて収集した各深度画像を2.5D点群に変換した。Kinectで計測した点群に遠い背景や近すぎるノイズ点が存在するので、フィルタをかけて2m以外の点を外れ値として除外した。収集したデータが高いクオリティのコンテンツを生成するモデルの訓練に利用されることを想定し、2.5D点群と3Dモデルを128x128x128の高解像度のボクセルにサンプリングした。

図1:前処理の流れ

4. ニューラルネットワークの構造

本研究で設計した敵対的生成ネットワークのアーキテクチャを図2に示す。生成器は2次元においてよく使用されるU-Net[2]を参考し、3Dデータに特化させたEncoder-Decoderモデルであり、識別器は3D畳み込みニューラルネットワークの2値分類器である。

Automatic Generation of 3D Content from a Single Depth Image Based on Deep Learning

†YANG, Menglong (myang@nagao.nuie.nagoya-u.ac.jp)
†NAGAO, Katashi (nagao@nuie.nagoya-u.ac.jp)
†Graduate School of Informatics, Nagoya University

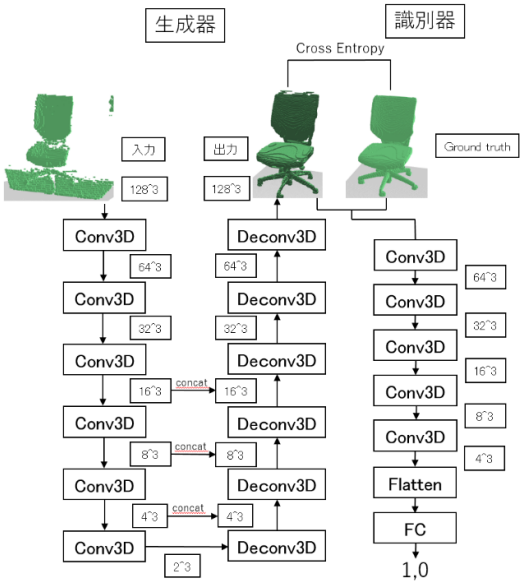

図 2:ネットワークのアーキテクチャ

U-Net では同じ次元である Convolution 層と Deconvolution 層が連結させているが、本研究では入力と出力との直な位置関連がないので、グローバルな特徴を掴む深い層のみを連結させた。入力された 2.5D ボクセルが、Convolution 層によってダウンサンプリングされ、Deconvolution 層によってアップサンプリングされていく。最後に入力と同じ次元の出力が求められ、各ボクセルにかけられた Sigmoid 関数によって点の存在が推定される。

識別器においては、一連の Convolution 層によって入力のボクセルからハイレベルな特徴が抽出され、Flatten 層に通してベクトルに展開される。全連結層に経由し Sigmoid 関数により、生成された物であるかどうかが推定される。これにより、生成器のパラメータは生成された物と真値との交差エントロピー誤差からだけでなく、識別器から逆伝播した勾配によっても更新される。

5. 評価実験

学習モデルを検証するために、収集したデータを以下のように 3 つに分類した。まず、5 種類の椅子のデータを、訓練されたモデルが訓練データにない種類の椅子への対応を検証するテストデータ(種類テストデータセット)とした。残り 25 種類の椅子に対し、1/3 の深度画像を、訓練データにない角度の深度画像への対応を検証するテストデータ(角度テストデータセット)とした。残りのデータを訓練データとした。

4 章で述べた生成器と同じ仕組みの Encoder-Decoder モデルをベースラインとして、テストデータにおいて生成されたコンテンツと真値とのボクセル IoU(intersection-over-union)の平均(大きいほど良い)と L1 距離の平均(小さいほど良い)を指標として敵対的生成モデルを比較し、結果は表 1 のようになった。各テストデータセットにおいて生成されたコンテンツの例を図 3 と図 4 に示す。本研究で提案した手法の方が各テストデータセットにおいて Encoder-Decoder モデルより質の高い 3D コンテンツが生成できることが確認された。

表 1 :テストデータセットにおける評価結果

モデル	IoU	L1 距離
Encoder-Decoder(角度)	0.5370	0.0094
敵対的生成モデル(角度)	0.5758	0.0086
Encoder-Decoder(種類)	0.2854	0.0191
敵対的生成モデル(種類)	0.3523	0.0156

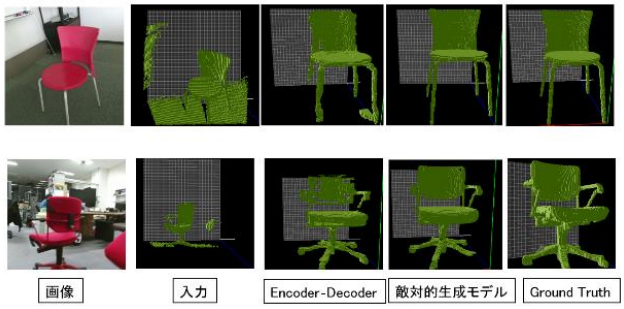

図 3:訓練データにない角度からの生成例

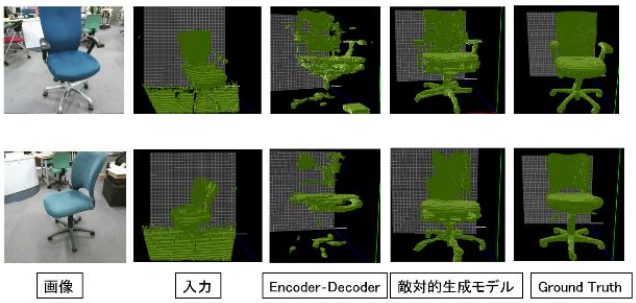

図 4:訓練データにない種類からの生成例

6. まとめ

本研究では、実世界のデータを取集し、敵対的生成ネットワークを用い、一枚の深度画像から 3D コンテンツを自動生成する手法を提案した。今後の課題として、更なる多様性への対応のために、実世界データと CAD モデルからの合成データを組み合わせてモデルを訓練することが挙げられる。

参考文献

[1] Goodfellow, Ian, et al. "Generative Adversarial Nets." Advances in Neural Information Processing Systems. 2014.

[2] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional Networks for Biomedical Image Segmentation." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2015.