U-Net を用いた単一画像の奥行き推定

北見工業大学　○八重樫亮汰，岩館健司，鈴木育男，渡辺美知子

要　旨

平面画像から 3 次元空間を予測することは，コンピュータビジョンの分野における主要な課題のひとつである．3 次元空間を予測する際，RGB 画像に加え，深度情報がよく用いられる．深度を計測する手段として，何らかのセンサやカメラを複数台用いる等の前提条件を要していた．本研究では，そのような画像外の情報を含まない，単一の RGB 画像からの深度推定を U-Net と呼ばれる畳み込みニューラルネットワークを用いて行う手法について提案し，その有効性について検証を行う．

1 はじめに

平面画像から 3 次元空間を推定することは，コンピュータビジョンの分野における主要な課題のひとつである．3 次元空間において奥行を推定する際，RGB 画像の情報に加え深度情報が用いられる．深度を計測する手段としては，計測対象に光や赤外線といった作用を与えてその変動を観測する方法や，複数台のカメラを用いて撮影されたステレオ画像から対応する点を導き，その「ずれ」を利用して推測する方法等がある．しかし，いずれの方法も奥行き推定の際，画像以外の情報を要する点で煩わしさが残る．

本研究では単一画像の情報から奥行き推定することを目的とし，U-Net による畳み込みニューラルネットワークを奥行推定に用いる手法を提案する．本論文では，提案手法の有効性についてテストデータによる検証を行う．

2 提案手法

2.1 Convolutional Neural Network (CNN)

本研究では，奥行推定に畳み込みニューラルネットワーク（Convolutional Neural Network．以下 CNN）を用いる．CNN は，生物の視覚野に基づいた構造を持った順伝播型ニューラルネットワークの一種である．従来の順伝播型ニューラルネットワークは，畳み込み層とプーリング層という独特の構造により構成される点で異なる．

畳み込み処理では，(1)式により入力画像に存在する各フィルタが表す特徴を強調することができる．

$$a_{ij} = \sum_{p=0}^{H-1} \sum_{q=0}^{H-1} x_{i+p,j+q} h_{pq} \quad (1)$$

ここで，xは入力画像，hは特徴フィルタを表す．

一方，プーリング処理は畳み込み処理の後に行われ，抽出された特徴の位置感度を低下させる働きを持つ．プーリング処理では，入力画像の任意の正方領域内での最大値を選ぶ最大プーリング（max pooling）がよく使用され，本稿においても最大プーリングを行う．このプーリング処理により，入力値の次元削減が可能となる．

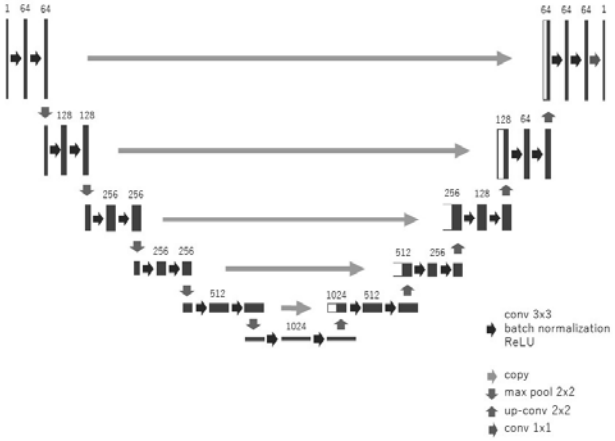

図 1．ニューラルネットワークの構成

2.2 ニューラルネットワークの構成

本研究で用いるネットワークの構成，及び各パラメータの設定を図 1 に示す．これらの設定は，Olaf[1]らによって考案された U-Net を参考にしている．U-Net は細胞のセグメンテーションなどで成果を出しており，U-Net を用いることで画像にある物体の輪郭をはっきりとさせたまま奥行き推定ができるのではないかと推測した．

2.3 誤差の算出

推論画像を教師画像に近づけていくための誤差の算出には以下の平均二乗誤差を用いる．

$$\mathrm{L}(y, y^*) = \frac{1}{n}\sum_i (y_i - y_i^*) \quad (2)$$

ここで，yは推定値，y^*は教師，iは画素，nは画素数を表す．

3 検証実験

3.1 実験設定

提案手法により奥行推定が可能かどうか検証実験を行った．CNN の学習には，Microsoft Kinect の RGB カメラと深度カメラの両方で撮影された様々な屋内の画像で構成された NYU Depth Dataset V2[2]のデータセットを用いた．各画像サイズは 480×640 であるが，ネットワークへ入力する際には 96×128 にリサイズする．学習において，画像の左右反転と色相変換，クロップをランダムで行うことにより学習データの補強を行い，精度の向上を図った．学習に使用した各種パラメータを表 1 に示す．

3.2 精度の算出

推定結果の精度を測る尺度として，本稿では式(3)により計測する．

$$\max\left(\frac{y_i^*}{y_i}, \frac{y_i}{y_i^*}\right) = \delta < t(t \in [1.25, 1.25^2, 1.25^3]) \quad (3)$$

これは，CNN からの推定画像と教師画像の 1 ピクセルごとの差を計算し，差の閾値以下のピクセル数を正解として求めている．最終的に，画像サイズに対する正解ピクセルの割合を推定精度とした．

3.2 結果と考察

CNN の学習時における誤差と精度の遷移を図 2 に示す．Epoch が進むごとに誤差は収束しており，精度も向上していることから過学習を起こしている様子はなく，学習が順調に進んでいることがわかる．ただし，50 [epoch] 時点で精度の向上が停滞している状況である．

また，推論結果の一例を図 3 に示す．得られた推定画像からは，輪郭はぼやけているが深度の勾配については獲得されていることがわかる．一方で，カメラから近くにある物体の深度については正確に推定できていない．この原因については，カメラの近くにある学習データが少なかったため偏った学習が行われたのではないかと考えられる．

4 おわりに

本研究では，U-Net を用いた単一画像からの奥行き推定を目的とした．検証実験から，奥行深度の勾配についてはおおよそ推定されているが，近距離の推定に課題があることが分かった．

今後は，精度が出づらい深度を調べ，改善点を探っていきたい．また，深度の差から精度を算出するなど精度の算出方法についても再検討していく．

表 1．学習パラメータ設定

バッチ数	8
Epoch 数	200
学習率	0.00002
L2 正則化	0.001
Optimizer	Adam

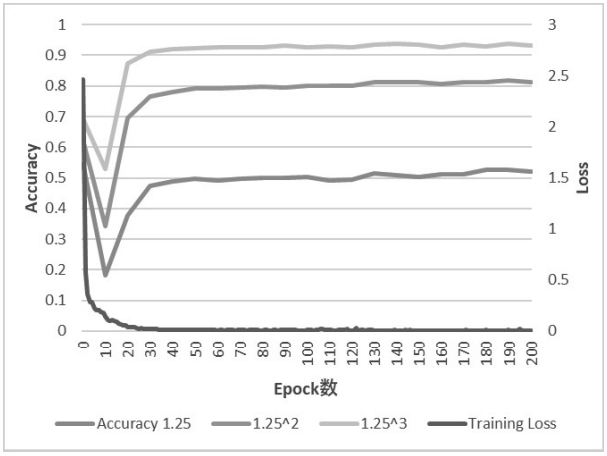

図 2．誤差の遷移

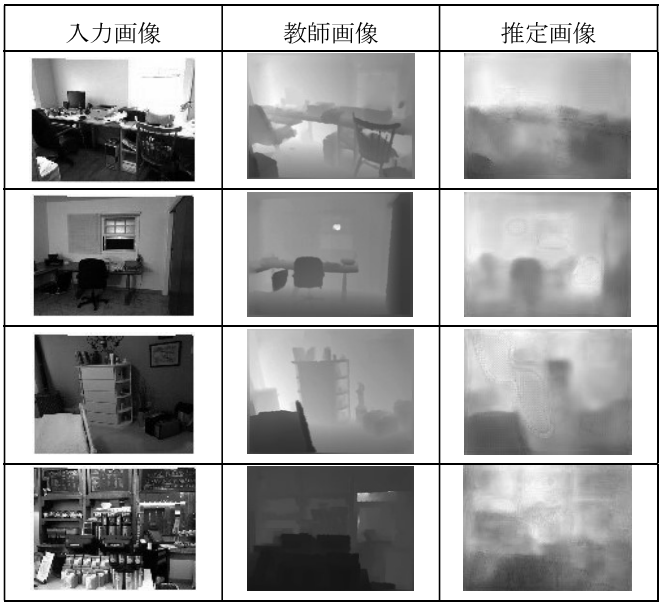

図 3．推論結果

参考文献

[1] Olaf Ronneberger, Philipp Fischer, Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”

[2] Nathan Silberman, Pushmeet Kohli, Derek Hoiem, Rob Fergus. “Indoor Segmentation and Support Inference from RGBD Images” ECCV 2012