単眼深度推定における物体認識の有効性

AZUMA, Fukuo / 吾妻, 福央

(出版者 / Publisher)
法政大学大学院理工学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 理工学研究科編
(巻 / Volume)
63
(開始ページ / Start Page)
1
(終了ページ / End Page)
4
(発行年 / Year)
2022-03-24
(URL)
https://doi.org/10.15002/00025369

単眼深度推定における物体認識の有効性

EFFECTIVENESS OF OBJECT RECOGNITION IN MONOCULAR DEPTH ESTIMATION

吾妻福央
Fukuo AZUMA
指導教員　平原誠

法政大学大学院理工学研究科応用情報工学専攻修士課程

In general, existing monocular depth estimation methods use only images from a single camera as input data, particularly RGB images. In this study, therefore, we focus on the retinal image size of known objects, which is an important depth perception cue in human vision. We investigate whether adding object recognition results (segmented images) to the input of CNN models improves the accuracy of depth estimation.

Key Words *: Depth estimation, Object recognition, CNN, Human depth perception*

1.　はじめに

単眼深度推定とは単眼カメラからの情報を基に深度(奥行き)を推定する技術のことで，主にRGB画像から深度を推定する[1]．本研究では単眼深度推定においてRGB画像の他に物体認識画像をCNNに追加入力して深度を推定することでその有効性を確認する．

2.　物体認識画像の追加意義

人間の単眼での奥行き知覚の手がかりの１つとして像（網膜像）の大きさがある．我々は物体の実際の大きさとその像の大きさから物体までの深度を予測することが可能である．このように物体の実際の大きさとその像の大きさを認識させるために，本研究では物体の種類ごとの物体認識画像をCNNに追加入力し深度推定精度の向上を図る．

3.　従来モデルと提案モデル

以前我々はRGB画像のみをCNNに入力し，深度画像を出力するエンコーダデコーダ型のモデルを作成した[2]．図１に従来モデルを示す．それに対して，今回提案するモデルを図２に示す．図２の提案モデルでは図１のモデルに物体認識画像を追加入力するようになっている．これら従来モデルと提案モデルの深度予測の精度を比較することによって物体認識画像をCNNに追加入力する有効性を確認する．

中間層の畳み込み層の活性化関数にはReLU関数，出力層の活性化関数にはSigmoid関数を用いる．誤差関数には誤差二乗和，フィルタの最適化にはAdamを用いる．学習にはミニバッチ学習を用い，ミニバッチ数は８とする．

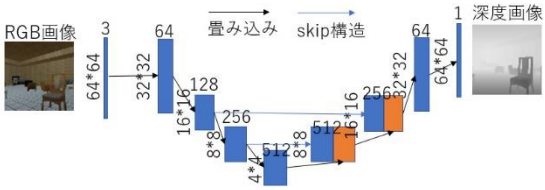


図１　従来モデル[2]

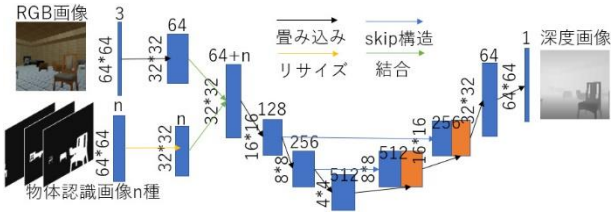
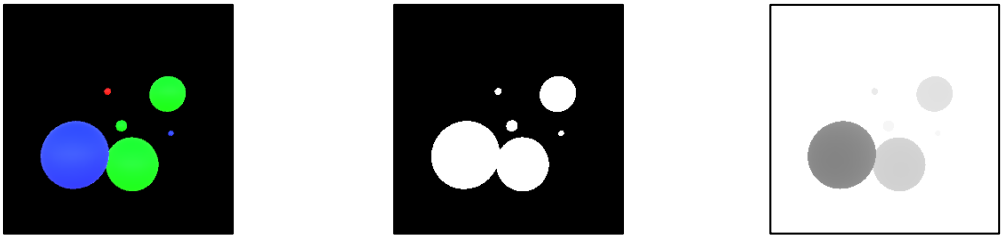
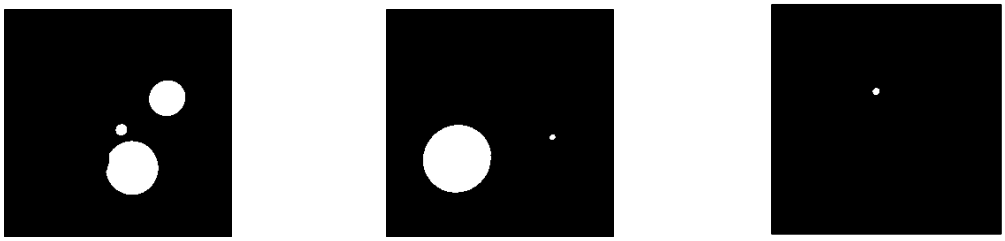


図２　提案モデル



(a)RGB 画像(物体確認用)　(b)二値画像　(c)理想深度画像



(d)物体認識画像(大)(3 個)　(e)物体認識画像(中)(2 個)　(f)物体認識画像 (小)(1 個)

図 3　本実験のデータセットの一例(出現パターン②)

４．検証実験

４．１．検証実験のデータセット

物体認識が可能になることによって深度推定の精度が上がることを確かめるために，検証実験では従来モデルによって物体を識別することが不可能なデータセットを使用する．方法としては，色も形も同じで，大きさだけ異なる 3 種類の球体オブジェクトがランダムな場所に出現する仮想空間を Unity 上で作成した．それぞれの球体の大きさは 2 倍ずつ半径が大きくなっている．なお，物体の出現パターンを以下の 2 種類用意した．

出現パターン①:球体オブジェクト大・中・小のそれぞれが一つだけ出現.

出現パターン②:球体オブジェクト大・中・小のそれぞれが 0 から 5 の間の一様乱数で決められた個数だけ出現.

これら 2 種類の物体出現パターンのデータセットを 602 セットずつ用意した．1 セットの内容は単眼カメラから得られる画像として二値画像，各大きさの球体に対応する物体認識画像，教師深度画像である．データセットの内訳として，602 セットのうち 416 セットを学習データに，176 セットを検証データに，10 セットをテストデータに使用する．

出現パターン②のデータセットの一例を図 3 に示す．図 3(a)の RGB 画像は我々人間が視覚的に球体オブジェクトを識別できるようにするための物体確認用 RGB 画像であり，実験には一切使用しない．球体オブジェクト大・中・小はそれぞれ緑色，青色，赤色で表示している．図 3(b)の二値画像を見ると，球体オブジェクト大・中・小の計 6 個の球体が映っているが，それらの区別を二値画像のみで識別することはできない．それは形が全て同じであり，深度がばらばらであるためである．図 3(d),(e)および(f)の各物体認識画像より，球体オブジェクト大・中・小のそれぞれの出現個数は 3 個，2 個，1 個となっていることが分かる．物体認識画像を追加入力すると物体認識可能となる．このため，２つのモデルの性能比較は物体認識画像の入力の効果を如実に表すことになるだろう．

４．２．検証実験の結果

出現パターン①と②の場合のテストデータに対する画像 1 枚あたりの平均予測深度誤差(±1SD)をそれぞれ図 4(a)と(b)に示す．共に物体認識画像を入力に追加した提案モデルのほうが，テストデータに対する予測深度誤差が小さくなった．

次に出現パターン①と②の場合のテストデータに対する予測深度画像の一例をそれぞれ図 5(a)と(b)に示す．深度画像は深度が近いほど黒色，遠いほど白色になっている．図 5(a)の RGB 画像を見ると，画面左上に球体オブジェクト小(赤)，画面中央に球体オブジェクト大(緑)が同じサイズで映っている．理想深度画像をみると，球体オブジェクト小(赤)の深度が近いことが確認できる．二値画像のみを入力した従来モデルではこの 2 つの球体オブジェクトを同じ深度として予測しているのに対し，物体

認識画像を追加入力した提案モデルでは球体オブジェクト小(赤)と球体オブジェクト大(緑)の深度に差をつけて予測しており，理想深度画像により近い結果となった．

次に出現パターン②の結果である図 5(b)を確認する．画面に大きく映っている球体オブジェクト中(青)と球体オブジェクト大(緑)について，理想深度画像を見ると，球体オブジェクト中(青)が近い深度であることが確認できる．二値画像のみを入力した従来モデルではこの 2 つの球体オブジェクトを同じ深度として予測しているのに対し，物体認識画像を追加入力した提案モデルでは球体オブジェクト中(青)と球体オブジェクト大(緑)の深度に差をつけて予測しており，理想深度画像により近い結果となった．

二値画像のみを入力した従来モデルでは大きさの異なる球体を認識することができず，仕方なく二値画像に映っている球体の像の大きさに応じた深度を出力しているのに対して，物体認識画像を追加入力した提案モデルは，大きさの異なる球体を正しく認識し，それぞれの球体オブジェクトの物体認識画像内の像の大きさから深度を推定したことで，理想深度画像に近い結果になったのだと考えられる．

以上の結果より，物体認識画像を追加入力することで，深度推定の精度が向上することを確認できた．

５． 本実験

５． １．本実験のデータセット

本実験では，現実空間風に作成した部屋の VR データセットを用いる．部屋の中にはイスやテーブルなど 7 種類の家具を複数個自然と感じる配置で設置した．7 種類の家具に対する物体認識画像は Segnet[3]のような物体認識モデルによって作成できることを前提としており，今回は VR 空間で撮影した誤りのない物体認識画像を使用する．部屋の数は 11 部屋とし，1 部屋ごとに 200 セット撮影したため，計 2,200 セットのデータセットとなる． その内 10 部屋の 1,400 セットを学習データ，600 セットを検証データ，残り 1 部屋 200 セットをテストデータとした．なお，テストデータは学習データおよび検証データに存在する家具や壁紙，部屋の形を使用し，家具の配置だけを変えることで撮影した．

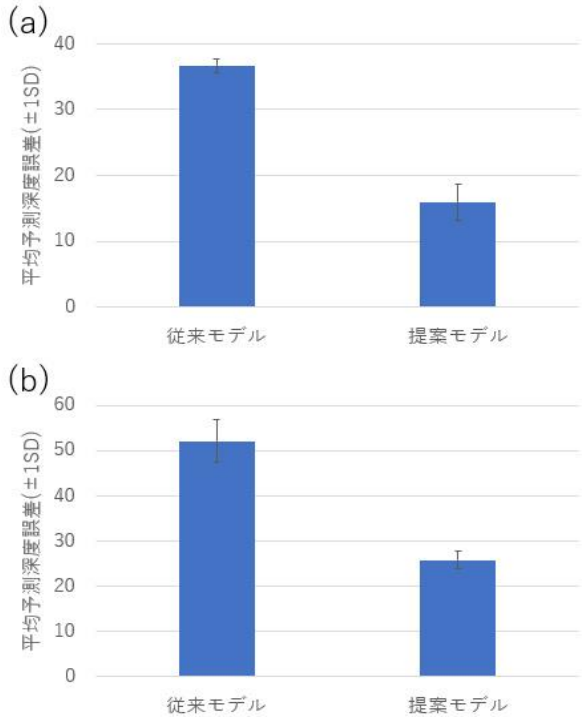


図 4　検証実験の平均予測深度誤差(±1SD)の比較，(a)出現パターン①，(b)出現パターン②

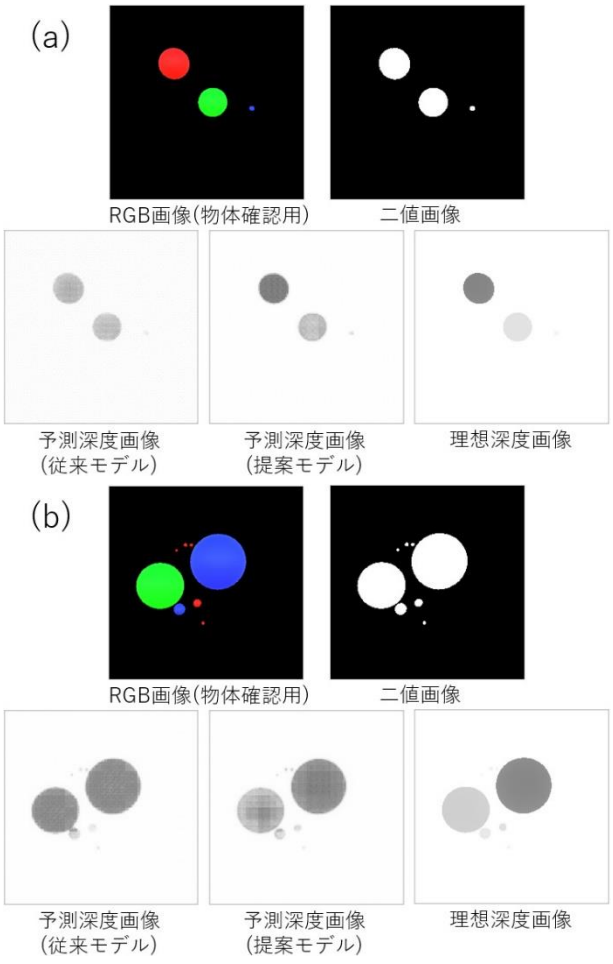


図 5　検証実験の予測深度画像の比較，(a)出現パターン①，(b)出現パターン②

５．２．本実験の結果

本実験のテストデータに対する画像 1 枚あたりの平均予測深度誤差(±1SD)を図 6 に示す. 従来モデルよりも物体認識画像を入力に追加した提案モデルのほうが，テストデータに対する予測深度誤差が小さくなった.

次に，本実験のテストデータに対する予測深度画像の一例を図 7 に示す．従来モデルの予測深度画像よりも提案モデルの予測深度画像の方が画面右手前のテーブルや画面左手前のベッドの深度が理想深度画像に近かった．それは物体認識画像の追加によって物体ごとの像の大きさを捉えるようになり，そこから深度を推定しやすくなったためだと考えられる.

次に従来モデルの予測深度画像は画面右手前のテーブル天板の上面と側面で深度に差をつけてしまっている．それに対し，提案モデルの予測深度画像では，テーブル天板の深度が適切に予測できている.RGB 画像のみを入力した従来モデルでは色情報だけにより物体を認識しており，光の当たり具合によって色が変わってしまったテーブル天板の上面と側面で別の物体だと判断してしまったのだと考えられる．それに対して提案モデルでは，物体認識画像を追加したことによってテーブルの形状を正確に把握することが可能となり，テーブル全体の深度を適切に予測できたと考えられる.

６．結論と今後の展望

CNN を用いた単眼深度推定において RGB 画像の入力だけでは物体の形状や深度を正確に把握できていない．その場合，物体認識画像の追加は有効な手法である.

現在，CNN に追加入力している物体認識画像は VR 空間で撮影した誤りのない画像を使用している．しかし実際に現実空間で本研究の提案モデルを適用する場合，正確な物体認識画像を用意する必要がある．例えば RGB 画像から物体認識画像を予測する SegNet[3]というモデルがある．そのようなモデルを使って精度の高い物体認識画像を用意できるなら，それらは CNN における単眼深度推定の入力として有効なものとなるだろう.

最後に現実空間には様々な物体が存在する．本研究の提案モデルでは,認識させる物体の種類が増えるほど CNN への入力画像が増えてしまうといった課題がある．モデルの規模削減が今後の課題である.

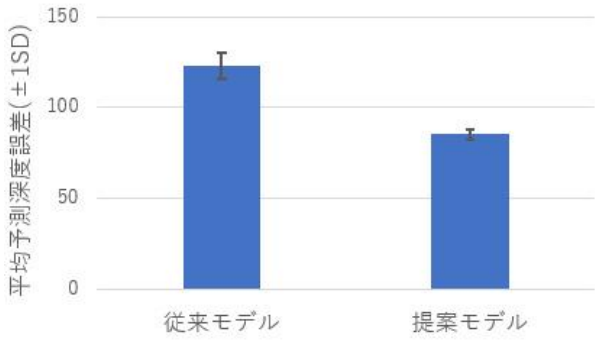


図 6　本実験の平均予測深度誤差(±1SD)の比較

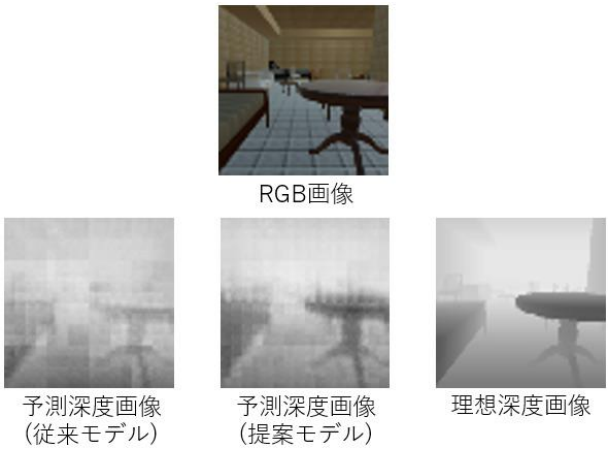


図 7　本実験の予測深度画像の比較

謝辞

本研究を進めるにあたり，ご指導いただいた修士論文指導教員平原誠准教授に心より感謝いたします.

参考文献

1）D.Eigen, et al.: Depth map prediction from a single image using a multi-scale deep network, NIPS14, Vol.2, pp.2366-2374, 2014.

2） 吾妻福央, 平原誠：物体認識結果を用いた深度推定, SIC2020-2, pp.15-18, 2020.

3）V.Badrinarayanan, et al.: SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol.39, No.12, pp.2481-2495, 2017.