深層学習による近距離で撮影された単眼画像からの深度推定

1215100　領内 あゆみ　【 知能情報学研究室 】

Depth Map Prediction from a Single Image in Close Distance using Deep Neural Network

1215100　RYONAI, Ayumi　【 Intelligent Informatics Laboratory 】

1　はじめに

深度推定とは，画像認識技術のひとつで，撮影するデバイスと被写体との距離の推定する技術のことである．距離を推定することで，画像に映っている空間の構造の把握することができる．正確な深度を推定することは，物体検出や画像のセグメンテーションに寄与する．

深度推定を行う際に入力として用いられる画像には，複眼画像と単眼画像の 2 種類がある．複眼画像は，2 枚以上の画像から 1 つのシーンの深度を計算する．この場合，人間が両眼の視差から物体の立体感を感じ取るように，それぞれの画像の差分から画像中の物体までの距離を推定する．一方，単眼画像の場合は 1 枚の画像から推定を行う．この際にデータとして用いられるのは，自動車の運転席からの画像や景色画像が主である．従来の研究では，中〜遠距離の画像がデータとして用いられていて，屋外に存在する物体間，あるいは部屋の中の家具間等の位置関係の把握が主な目的で，各物体の表面は滑らかなものとみなし，その表面上の細かな凹凸は対象として来なかったため，1m 以内の近距離を対象として深度推定を行った研究はない．

そこで，本研究では，低コスト組み込み機器向けに，距離計測機器を用いず深層学習による，単眼カメラ近距離カラー画像からの深度推定システムを構築する．

2　CNN による単眼深度推定

深度推定を行うために用いられている手法は主に 3 つに分けられる．1 つ目は固定焦点距離画像を用いる手法で，幾何学モデルに基づいたもの [1] であり，矩形の組み合わせで室内の様子を表すが，特定の構造のシーンにのみ適用が可能という制限があった．2 つ目は，ノンパラメトリックな手法 [2] であり，意味的に似通った外観であれば，類似の深度分布を有するという仮定が前提である．3 つ目が，近年のディープラーニングを用いたもので，上記のような制約にとらわれずにより一般的な画像に対して深度推定を行うことが可能である．

本研究では，Eigen らの提案した深度推定のための深層学習の手法 [3] を応用し，これを近距離単眼画像に適用する．

3　深度のファジィ論理表現

本研究で用いる深度カメラ (Intel Realsense Depth Camera D435) は，アクティブ赤外線光を用いるステレオビジョンの深度カメラ であり，有効距離は 0.2-10m，深度は 1mm 単位で取得できる．得られるデータは 16 ビットである．これを 10cm 程度の厚さの差異のみの近接画像にそのまま用いると，16 ビットのうち 7 ビット程度の情報しか用いられず，これを画像化すると深度の差が見られない一様な画像になる．画像のダイナミックレンジの変換として，ガンマ補正やより複雑なトーンカーブによる非線形変換が良く用いられるが，パラメータの設定に統一的な方法はなく，本研究では有効レンジが最大輝度の 1% 以下のような場合は設定が困難である．

そこで，本研究では各画素の深度を「厚さ」のファジィ論理値として表現する．本研究では，実験の対象物であるカットニンジンがベルトコンベア上に流れるものとして，カメラをコンベアの直上 420mm に設置されている．ニンジンの厚さのファジィメンバシップ関数として，深度値 $[d_{\mathrm{near}}, d_{\mathrm{far}}]$ をファジィメンバシップ値 $[0, 1]$ とする台形型を採用する (図 1)．本研究では，$d_{\mathrm{near}} = 350, d_{\mathrm{far}} = 450$ としている．

ファジィ化された深度画像を，文献 [3] のネットワークに入力することで，RGB 画像から深度を推定するためのパラメータの学習を行う．撮影したニンジンの画像

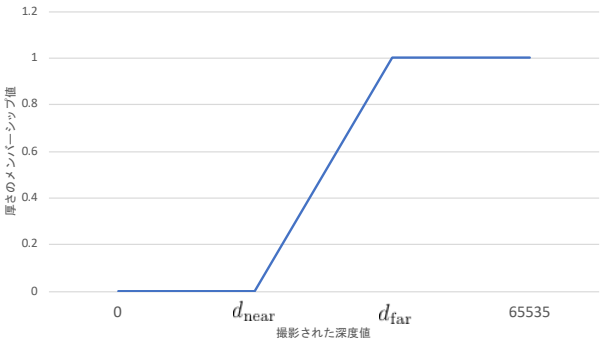

図 1　厚さを表す台形ファジイメンバシップ関数

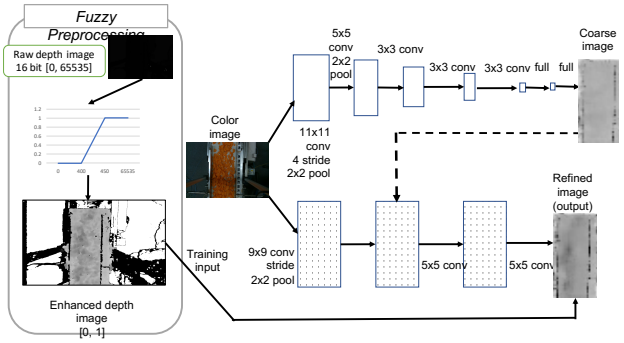

図 2　ネットワーク構造

表 1　実験結果

PSNR	SSIM
9.049	0.5985

のうち，4295 枚の RGB 画像と深度画像をネットワークへの入力とし，学習を行った．

3.1　ネットワーク構造

使用した DNN は，図 2 のような畳み込みネットワークの 2 段構成をとっている．1 段目の Coarse ネットワークと 2 段目の Refine ネットワークを，1 段ずつ学習に用いる．まず Coarse ネットワークが大域的に画像の深度を予測する．次に，入力画像と Coarse ネットワークの出力を基にして，Refine ネットワークが局所領域ごとに，より精度の高い予測を行う．ネットワーク全体としては，入力として 1 枚の画像をとり，画素単位で深度を推定した画像を出力する．

4　評価と考察

4.1　定量的評価

学習用に用いた画像データと重複しない RGB 画像 3,157 枚を用いた．推定された深度画像を，PSNR（ピーク信号対雑音比）および SSIM（Structural SIMilarity）の 2 つの指標を用いて評価した．表 1 がその結果である．PSNR は画素ごとに実際に撮影された深度画像（ground truth）との類似度を評価する．数値が高いほど，推定精度が高いことを表す．一方 SSIM は画素ごとだけではなく，周囲の画素との相関まで算出に用いている類似度評価指標である．このため，PSNR に対して人間の感覚に近い類似性を求めることができる．評価の際には両画像の解像度が同一である必要があるため，推定深度画像を拡大し，撮影深度画像の解像度に合わせてから計算を行った．

また，学習経過中に推定画像と ground truth の相違度の推移を表したものが図 3 である．この指標には MSE（平均二乗誤差）を用いた．横軸はエポック数，縦軸は MSE を表す．

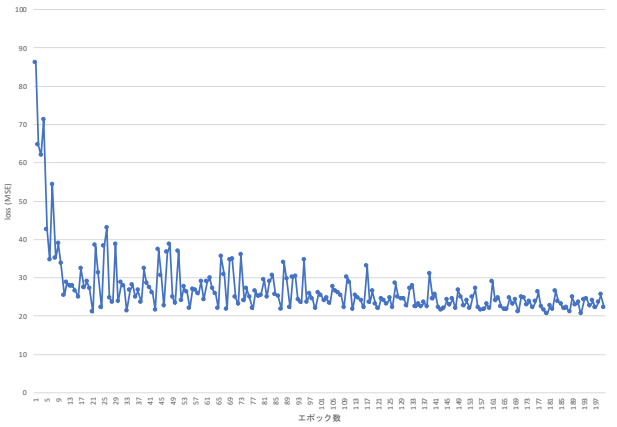

図 3　学習の進捗に伴う loss 値の変化

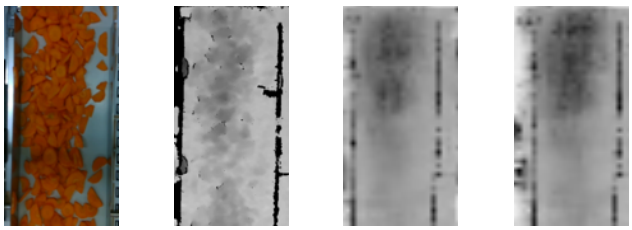

図 4　左から RGB 画像，撮影深度画像，深度推定画像 (80epoch)，深度推定画像 (137epoch)

4.2　定性的評価

図 4 は，それぞれ学習の入力に用いたカラー画像，それに対応する深度カメラで撮影された画像および 80 epoch, 137 epoch 時点のパラメータを基に推定された深度画像である．結果の画像は，ベルト上の対象物の厚みをおおよそ推定できていることを示していると考える．

5　まとめ

深度カメラ画像を教師信号として，Eigen の CNN 深度推定モデルを用いて単眼カラー画像からの深度推定を行った．50cm 以内の近距離を対象とするため，そのまま深度画像を用いても効率的な学習が行えないため，「厚さ」を表現するファジィメンバシップ値に変換して教師データとした結果，良好な深度マップが得られた．

参考文献

[1] Hedau V, et al., “Thinking inside the box: Using appearance models and context based on room geometry,” ECCV 2010, pp. 224–237.

[2] K. Karsch, et al., “Depth transfer: Depth extraction from video using non-parametric sampling,” IEEE T. PAMI, 36(11), pp. 2144–2158, 2014.

[3] Eigen D, et al., “Depth map prediction from a single image using a multi-scale deep network,” NIPS 2014, pp. 2366–2374.