I-020

単眼深度推定を用いた三次元映像生成に関する基礎検討

A Basic Study on View Synthesis Using Monocular Depth Estimations

杉江 孝士 *　都竹 千尋 *　高橋 桂太 *　藤井 俊彰 *
Takashi Sugie　Chihiro Tsutake　Keita Takahashi　Toshiaki Fujii

1　はじめに

近年，任意の視点から対象を視聴できる三次元映像技術が注目を集めている．カメラ幾何に基づく三次元映像生成では，多視点画像からカメラの位置・姿勢を推定できると同時に，カメラから三次元物体までの距離も推定できるため，これらの情報を活用して三次元映像を生成できる．この方法は，カメラの回転や平行移動といった明示的な運動モデルに基づいており，対象物体とカメラの関係を外在的に対応づけるのが特徴である．しかし，シーンにオクルージョンなどが含まれる場合，距離の推定結果が曖昧になり，三次元映像に視覚的なノイズが生じる．

この問題を解決するために，ここ数年で急速に発展した単眼深度推定技術の相補的活用を提案する．単眼深度推定は，画像に内在する暗黙的な単眼手がかりを利用することで，単一のカメラのみから物体までの距離推定を可能とする技術であり，カメラ幾何的には必ずしも正しいとは限らないが，視覚的には矛盾の少ない画素単位の距離を与える．我々の目的は，この単眼深度推定によって暗黙的に得られる深度情報を，カメラ幾何によって明示的に得られる深度情報と融合することで，視覚的に自然な三次元映像を生成することである．本稿はその基礎検討であり，ステレオ幾何，および深層学習に基づく単眼深度推定器 MiDaS [1] に焦点を絞る．まず，MiDaS で得られる単眼視差を用いて，オクルージョンなどが原因で欠損した両眼視差を復元する手法を提案する．また，復元した両眼視差を用いて実際に三次元映像を生成し，視覚的に評価することで，提案手法の有効性を確認する．

2　関連研究

2.1　ステレオ幾何と三次元映像

ステレオ幾何では，二台のカメラで撮影された画像を用いて，三角測量の原理によって対象物体までの距離をパッシブに推定する．ここでは，平行ステレオを仮定する．カメラのベースラインを L，焦点距離を f，画像座標を (x, y)，左画像を基準とした両眼視差を $d^{\mathrm{stereo}}_{x,y}$ として，左カメラの光学中心をワールド座標系の原点とすると，三次元物体 $P(X, Y, Z)$ の座標は次式で表される．

$$X = \frac{Lx}{d^{\mathrm{stereo}}_{x,y}},\ Y = \frac{Ly}{d^{\mathrm{stereo}}_{x,y}},\ Z = \frac{Lf}{d^{\mathrm{stereo}}_{x,y}} \tag{1}$$

従って，d^{stereo} が既知であれば，式 (1) に基づいて物体の三次元座標の復元が可能となる．また，モデルベースドレンダリングやイメージベースドレンダリングによって，復元した三次元座標から三次元映像を生成できる．ただし，ステレオ画像にオクルージョンが含まれる場合や，物体表面のテクスチャ不足によって画素の対応関係が取れない場合に，d^{stereo} を求めることが不可能である．そのような領域に対して，d^{stereo} の内挿が考えられるが，三次元映像に視覚的なノイズが生じることが知られている [2]．

2.2　MiDaS（Mixing Datasets）[1]

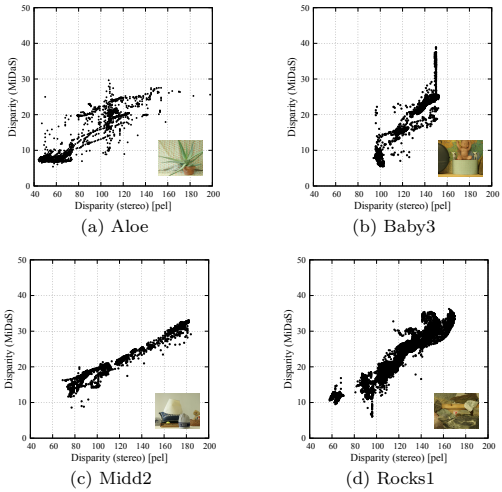
一般に，研究用途に公開されている深度情報は，アクティブ計測とパッシブ計測によって得られたものが混在しており，機械学習に基づく単眼深度推定においては，単位（スケールおよびオフセット）が統一されていない多様な深度情報を扱うことは極めて重いタスクとなる．この問題を解決するために，MiDaS では正解深度情報を視差 d^{gt} に変換する．また，後述するスケール a_1, a_2 とオフセット b_1, b_2 を用いて，損失関数

$$\mathrm{Loss} = \|\hat{d}(\theta) - \hat{d}^{\mathrm{gt}}\|_1 + \alpha\|\nabla(\hat{d}(\theta) - \hat{d}^{\mathrm{gt}})\|_1,\ \alpha > 0 \tag{2}$$

$$\hat{d}(\theta) = a_1 \cdot d(\theta) + b_1,\quad \hat{d}^{\mathrm{gt}} = a_2 \cdot d^{\mathrm{gt}} + b_2 \tag{3}$$

を定義しており，式 (2) を最小化する推定器を学習することで，単一の画像（入力）から単眼視差 d^{mono} を出力する．ここで，θ はネットワークパラメータである[*1]．また，a_1, a_2 と b_1, b_2 は各々 $d(\theta)$, d^{gt} のスケールとオフセットであり，データからロバスト推定される．

MiDaS は，オクルージョンやテクスチャの有無に依存せずに画素単位の視差を推論できる点が，ステレオ幾何とは根本的に異なる．従って，ステレオ法における欠損視差を MiDaS の視差で補完することによって，画素単位の両眼視差の復元が可能となり，三次元映像生成の高精度化が望める．ただし，MiDaS が推論する d^{mono} は式 (3) による補正が施されており，d^{stereo} と d^{mono} は単位が異なるため，これを逆補正した上で補完する必要がある．



(a) Aloe　(b) Baby3　(c) Midd2　(d) Rocks1

図 1　両眼視差と単眼視差の関係

3　提案手法

文献 [3] のステレオデータセット Aloe，Baby3，Midd2，Rocks1 に対して，左画像を基準とした両眼視差 d^{stereo} と MiDaS の単眼視差 d^{mono} の散布図を図 1 に示す．単眼視差には推定誤差が含まれているが，単眼視差と両眼視差は概ねアフィンの関係 $d^{\mathrm{mono}} = a \cdot d^{\mathrm{stereo}} + b$ にあると言える．これは，MiDaS で式 (3) のように線形モデリングが用いられることから，妥当な結果であると言える．提案手法では，両眼視差の欠損領域を Ω，その補集合を Ω^c として，単眼視差と両眼視差を関係付ける a, b を Ω^c 上で求める．具体的には，最小二乗問題

$$a, b = \underset{a', b'}{\mathrm{argmin}} \|a' \cdot d^{\mathrm{stereo}}_{\Omega^c} + b' - d^{\mathrm{mono}}_{\Omega^c}\|^2 \tag{4}$$

を解く．また，単眼視差を $(d^{\mathrm{mono}}_{\Omega^c} - b)/a$ と逆補正することで，$d^{\mathrm{stereo}}_{\Omega^c}$ との単位が揃う．

次に，$(d^{\mathrm{mono}}_{\Omega} - b)/a$ を用いて欠損視差 $d^{\mathrm{stereo}}_{\Omega}$ を補完し，映像生成に用いる視差 d を得る．次章の実験で示すように，$(d^{\mathrm{mono}}_{\Omega} - b)/a$ を

* 名古屋大学 大学院工学研究科 情報・通信工学専攻

[*1] https://github.com/intel-isl/MiDaS で配布されている学習済みのモデル (dpt-large) を使用した．

d_Ω^{stereo} に単純にはめ込むと境界 $\partial\Omega$ で視差が不連続となり，生成した三次元映像に視覚的なノイズが生じる．そこで，Ω 上で両者を違和感なく合成する．これは，いわゆる画像のシームレス合成問題に帰着できるため，提案手法では代表的な画像合成法である Poisson blending [4] に着目し，これを視差合成に応用する．具体的には，Ω 上の勾配マッチング問題

$$\min_d \int \left|\nabla d_\Omega - a^{-1}\nabla(d_\Omega^{\text{mono}} - b)\right|^2 \text{ s.t. } d_{\partial\Omega} = d_{\partial\Omega}^{\text{stereo}} \tag{5}$$

から導出されるポアソン方程式

$$\Delta d_\Omega = a^{-1}\Delta d_\Omega^{\text{mono}} \text{ s.t. } d_{\partial\Omega} = d_{\partial\Omega}^{\text{stereo}} \tag{6}$$

を解くことで，映像生成に用いる視差 d を得る．ここで，$\Delta(b/a) = 0$ であり，ポアソン方程式の解はスケール a には依存するが，オフセット b には依存しない．従って，$\partial\Omega$ における不連続性の源泉とも言える b の推定誤差は無視できるため，シームレスな視差合成が可能となる．

4　実験

ステレオデータセット Aloe と Midd2 に提案手法を適用し，単眼視差を用いて欠損状態の両眼視差を復元する．さらに，復元視差を用いて，三次元映像の 1 シーンである中央視点の画像（以下，単に中央画像）を生成し，視覚的な評価を行う．

一般に，ステレオ幾何における三次元映像生成では，視差推定によって両眼視差を得るが，ここでは左視点を基準とした正解視差が与えられているとする，なお，これにはオクルージョンなどに起因する欠損領域 Ω が含まれる．以後，この視差情報を Stereo と表記し，欠損領域には 0 を割り当てる．次に，左視点の画像を MiDaS に入力することで，単眼視差が得られる．Ω^c 上で単眼視差と両眼視差を対応付けることで a, b を求め，補正された単眼視差 $(d^{\text{mono}} - b)/a$ を得る．この補正済みの単眼視差を Mono と表記する．なお，Aloe と Midd2 の a は $0.193, 0.197$，b は $-2.51, -2.24$ であった．最後に，Stereo の欠損領域 Ω を，Mono を用いて補完する．Mono の視差を Ω に単純にはめ込んだ視差情報を Naive と表記する．一方，提案手法では，この補完処理に Poisson blending を用いる．実験では，これら 4 つの視差情報（Stereo，Mono，Naive，提案手法）の比較に加え，これらの視差情報を用いた映像生成の結果も比較する．映像生成では，左視点の画像と視差情報を用いて，forward warping により中央画像を生成した．

Aloe の各視点の画像，視差マップ，および生成した中央画像を各々図 2–4 に示す．図 3 の二段目は，同図の一段目に示す赤領域の拡大画像である．また，図 4 の一段目は生成した中央画像の拡大画像であり，図 2 の青領域に対応する．図 4 の二段目は正解画像（中央画像）と生成結果の誤差を可視化した画像である．Stereo では，アロエで隠蔽された壁紙の視差が欠損している（$d_\Omega^{\text{stereo}} = 0$）ため，中央画像の壁紙に大きなずれが生じる．Mono では，推定器は尤もらしい視差情報を出力する一方で，2 本のアロエが同じ視差であると誤推定しており，アロエの棘等の細部の情報が失われる．Naive では，視差の境界部分 $\partial\Omega$ が不連続であり，Stereo と同様に壁紙が歪むため，誤差が大きくかつ違和感がある．これらの結果に対し，提案手法では $\partial\Omega$ で復元視差が連続であり，画像の歪みも生じておらず，最も自然なレンダリング結果であると言える．よって，提案手法は視覚的な評価において有効であると結論づけられる．

また，Midd2 の各視点の画像，視差マップ，および生成した中央画像を各々図 5–7 に示す．Aloe の実験結果と同様に，提案手法のレンダリング結果が最も自然であることが確認できた．

5　まとめ

両眼視差と MiDaS によって推定される単眼視差をシームレスに合成する手法を提案し，三次元映像生成を行った．また，視差情報と生成結果を評価し，提案手法の有効性を視覚的に確認した．今後は，提案手法を多視点ステレオに拡張する予定である．

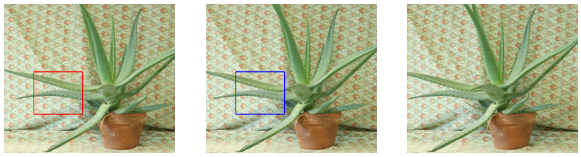
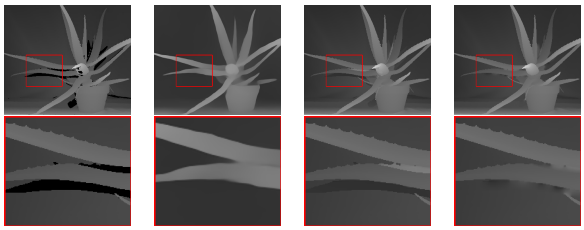
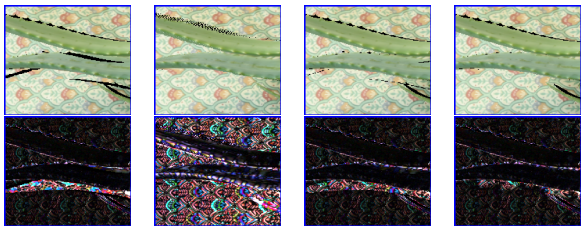


図 2　左画像，中央画像，右画像



(a) Stereo　(b) Mono　(c) Naive　(d) 提案手法

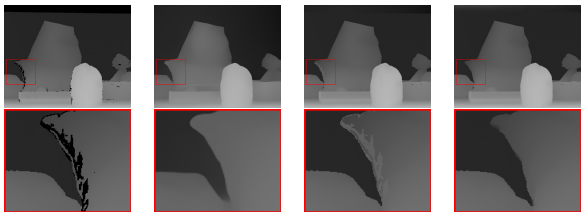
図 3　視差マップの比較



(a) Stereo　(b) Mono　(c) Naive　(d) 提案手法

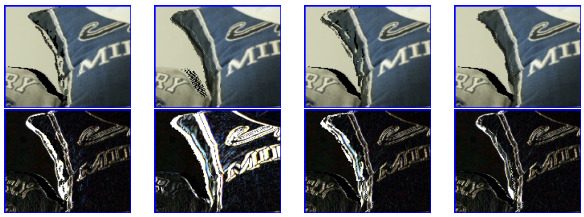
図 4　生成した中央画像の比較

図 5　左画像，中央画像，右画像



(a) Stereo　(b) Mono　(c) Naive　(d) 提案手法

図 6　視差マップの比較



(a) Stereo　(b) Mono　(c) Naive　(d) 提案手法

図 7　生成した中央画像の比較

参考文献

[1] R. Ranftl et al., “Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer,” *IEEE Trans. Pattern Anal. Mach. Intell.* 2020.

[2] M. Bleyer et al., “PatchMatch Stereo - stereo matching with slanted support windows,” *British Machine Vision Conference*, 2011.

[3] H. Hirschmuller and D. Scharstein, “Evaluation of cost functions for stereo matching,” *IEEE Comput. Soc. Conf. Comput. Vis. Pattern. Recognit. (CVPR)*, 2007.

[4] P. Pérez et al., “Poisson image editing,” *ACM Trans. Graph.*, pp. 313–318, 2003.