複数カメラを用いた単眼深度推定の融合による新規のステレオビジョン

○筑後 光（中央大学） 高木 大輔（中央大学） 野中 隼矢（中央大学）
Pathak Sarthak（中央大学） 梅田 和昇（中央大学）

本研究では，複数カメラを用いたステレオビジョンと単眼深度推定との融合手法を提案する．ステレオビジョンは幾何的に距離を計測するため，テクスチャのある領域では距離画像計測を行うことが出来るが，テクスチャの少ない領域では困難である．一方，単眼深度推定手法はテクスチャの多少にかかわらず距離画像計測を行うことが出来る．しかし，学習データに依存してしまうことや細かい特徴を計測することが困難である．そこで，距離を比較し，信頼度の高い距離を選択することで高精度な 3 次元計測を行うことを目指す．

1. 序論

1.1 研究背景

近年，自動運転に関する技術の開発が盛んに行われており，既に実用化されているシステムもある [1, 2]．自動運転システムでは，周囲の 3 次元情報を取得する必要がある．3 次元情報を取得するための代表的なセンサとして，ステレオカメラや LiDAR，ソナーなどがある．車の周辺環境を把握するために使用されている距離センサの中でも特にステレオカメラは，色情報を取得可能であることや低コストな LiDAR よりも高い計測密度で計測を行うことが可能である．

ステレオカメラでは，2 台以上のカメラを用いて視差を求め，三角測量の原理を用いることで 3 次元計測を行っている．複数の視点からの情報を用いることで幾何的に深度を求めることが出来るため，高精度かつ高密度な計測が可能となる．ステレオカメラを用いた 3 次元計測手法として，セミグローバルマッチングや RAFT などがある [3, 4]．しかし，これらの手法ではテクスチャの多少に影響を受けやすく，テクスチャレスなシーンでは 3 次元計測が困難である．単眼カメラから 3 次元計測を行う手法の 1 つとして，MiDaS や DenseDepth のような単眼深度推定手法が挙げられる [5, 6]．単眼深度推定手法は大量の画像データを学習することで単眼画像のみで 3 次元計測を可能にしている．また，テクスチャの多少にかかわらず 3 次元計測を行うことが可能である．しかし，学習データに依存するため，学習データと異なるシーンにおいて 3 次元計測を行うことが困難である．また，細かい特徴を計測することが出来ないといった課題もある．そこで本研究では，ステレオビジョンの RAFT と単眼深度推定手法の MiDaS から得られた距離を比較し，信頼度の高い距離を選択することで高精度な 3 次元計測手法を目指す．

1.2 関連研究

ステレオビジョンと単眼深度推定手法を融合する手法として，ルールベースで行う方法と深層学習ベースで行う方法がある．ルールベースで行う方法として，セマンティックセグメンテーションを用いて融合する手法が挙げられる [7]．しかし，誤差の大きい距離が選択されていた場合，補間処理を行う際にその距離も使用するため，誤差が拡大する可能性がある．深層学習ベースで行う方法として，単眼カメラ画像から異なる視点の画像を生成し，融合を行う手法やステレオ画像から得られる視差画像を教師データとして使用し，単眼カメラ画像からの距離推定を行う手法などが挙げられる [8, 9]．しかし，学習時の環境に依存するため，学習データとは違う環境に対して精度が保証されないという課題がある．また，処理負荷が大きいためリアルタイム性に課題がある．

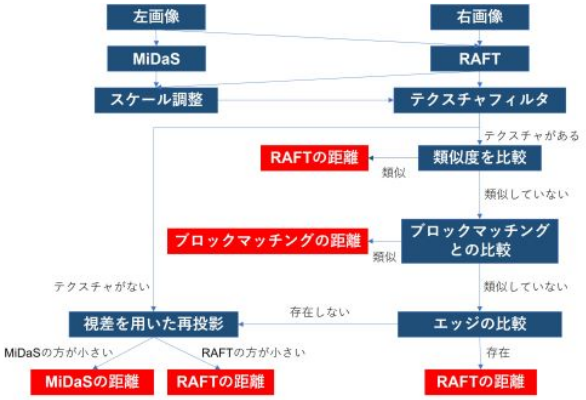


図 1 提案手法の流れ

2. 提案手法

2.1 概要

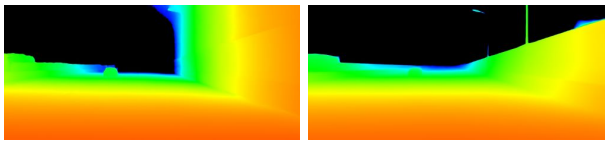
提案手法の流れを図 1に示す．まず，ステレオカメラから得られたステレオ画像より，MiDaS を用いた単眼深度推定と RAFT を用いたステレオビジョンを行う．次に RAFT で得られた距離と MiDaS で得られた相対距離を用いてスケール調整を行う．その次に MiDaS で得られたスケール調整後の距離と RAFT で得られた距離の融合を行う．融合過程では，類似度の高さやテクスチャの有無などを比較していくことで距離を選択する．

2.2 ステレオビジョン

本研究では，ステレオビジョン手法として，先述したように RAFT を用いる．RAFT は連続する 2 枚の画像を入力することで視差画像を出力する．本研究では，ステレオカメラで得られた左右の画像を入力とする．RAFT は Feature Extraction，Computing Visual similarity，terative Update の 3 つのステップにより視差推定を行っている．Feature Extraction では，畳み込みニューラルネットワークを利用することで入力画像から特徴量を抽出する．次に，Computing Visual similarity では，左右画像間の類似度の計算を行う．Iterative Update では，

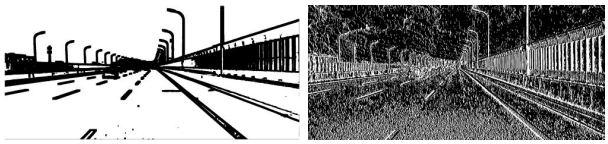
(a) 左画像　　(b) 右画像

図 2　入力画像



(a) RAFT　　(b) MiDaS

図 3　距離画像



(a) テクスチャフィルタ　　(b) ソーベルフィルタ

図 4　フィルタ処理の結果

視差推定を反復的に行う．RAFT に入力するステレオ画像とその結果を図 2，3(a) に示す．ここで，図 3(a) は距離が近いほど赤く，遠いほど青いことを表している．図 3(a) より，RAFT はステレオビジョン手法であるため，細かい特徴を計測出来ていることが分かる．

2.3　単眼深度推定

本研究では，単眼深度推定手法として，先述したように MiDaS を用いる．MiDaS は KITTI や NYUDepth v2 などの多様なデータセットを用いて学習されており，屋内外の様々な環境に対応することが出来る [10, 11].

MiDaS で得られる距離は相対距離である．そのため，カメラから計測物までの絶対距離を計測することが出来ない．そこで，RAFT で得られた距離を用いることでスケールを調整する必要がある．本研究では，テクスチャフィルタを用いたスケール調整を行う．テクスチャフィルタとは，距離算出を行わない領域を抽出する処理であり，ウィンドウ内の最大輝度差がしきい値未満の際に該当領域の距離算出を行わない．つまり，テクスチャの弱い部分を抽出することが出来る．図 2(a) にテクスチャフィルタを適用した結果を図 4(a) に示す．テクスチャフィルタが適用された領域を白，テクスチャフィルタが適用されなかった領域を黒で表している．テクスチャフィルタが適用されなかった領域はテクスチャが強いといえるため，RAFT で得られる距離の信頼度が高いと考えられる．MiDaS とのスケールを求める際，テクスチャフィルタが適用されなかった領域の各画素におけるスケールの平均とする．この処理によって得られたスケールを MiDaS の相対距離に掛けることで絶対距離への変換を行う．スケール調整の結果を図 3(b) に示す．ここで，図 3(b) は距離が近いほど赤く，遠いほど青いことを表している．

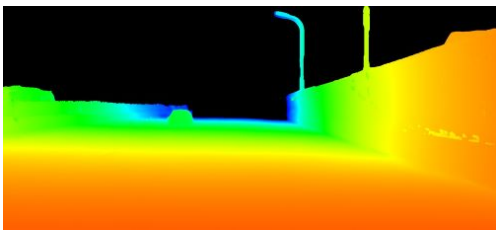


図 5　融合した距離画像

2.4　融合

2.4.1　類似度の計算

RAFT は幾何学的に視差を推定している.一方,MiDaS は深層学習を用いて相対距離を出力している．そのため，RAFT で得られた距離と MiDaS で得られた距離が類似している場合，RAFT で得られた距離の方が信頼度が高いと考えられる．また，実験的結果より多くの環境において RAFT で得られた距離の方が MiDaS で得られた距離よりも正確である．そこで，両手法で得られた距離の類似度が高い場合，RAFT で得られた距離を選択する．類似度が低い場合，後述する処理を用いて距離を選択する．本研究では，MAE（Mean Absolute Error）を用いて類似度を求める．環境によって，スケール調整を行う際に用いるスケールが大きく異なる．そのため，類似度を固定値にすることが出来ない．そこで，フレームごとに類似度を変化させることで対象フレームに適切な類似度を計算する．

2.4.2　ブロックマッチング

先述した類似度の処理より，類似度の低いピクセルに対してブロックマッチングを行う．ブロックマッチングによって得られた距離と RAFT と MiDaS で得られたそれぞれの距離の比較を行う．ただし，ブロックマッチングは輝度差に注目して行う手法であるため，テクスチャレスな領域においては誤マッチングが生じる可能性が高い．そのため，テクスチャフィルタが適用されなかったピクセルに対してのみブロックマッチングを行う．ブロックマッチングで得られた距離とどちらかの距離が類似している場合，ブロックマッチングで得られた距離を選択する．MiDaS は先述したように深層学習を用いている．また，RAFT は幾何学的に視差を推定しているが，手法の一部において．データセットを用いて学習を行っている．そのため，データセットの環境以外の場合，誤計測の可能性がある．一方，ブロックマッチングは非学習手法であるため，距離が類似している場合，ブロックマッチングで得られた距離が正確である可能性が高い．RAFT と MiDaS で得られたそれぞれの距離とも類似していない場合，誤マッチングの可能性があるためエッジを用いた処理を用いることで距離を選択する．

後述する再投影を用いて距離を選択する処理は処理負荷が大きい．処理負荷を低減させるため，エッジを用いた処理を行う．エッジがある場合，RAFT で得られた距離の信頼度は MiDaS で得られた距離の信頼度よりも高いと考えられる．そのため，テクスチャフィルタが適用された領域においてエッジがある場合，RAFT の距離を選択する．図 2(a) に示す入力画像にソーベルフィルタを適用した結果を図 4(b) に示す．エッジが存在する領域を白色で表している．

(a) 環境 1　(b) 環境 2

(c) 環境 3　(d) 環境 4

図 6　計測環境

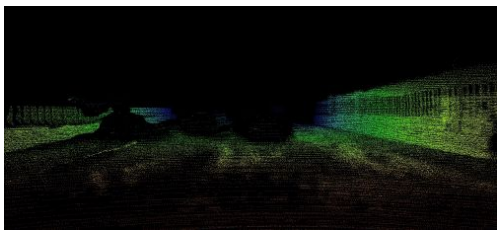


図 7　真値の距離画像

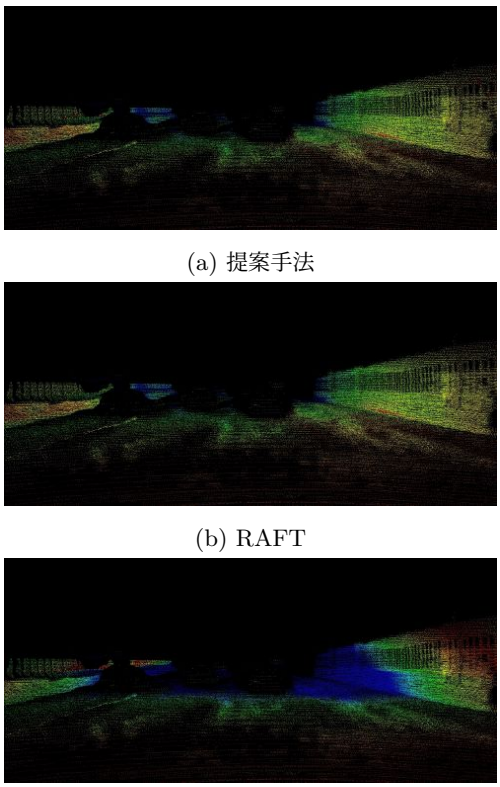
2.4.3　視差を用いた再投影

テクスチャフィルタが適用された領域やエッジがある領域はテクスチャが強いと考えられるため，RAFT で得られる距離の信頼度は高いと考えられる．しかし，テクスチャが弱い領域において，RAFT と MiDaS で得られる距離の信頼度は同等であると考えられる．そこで本研究では，視差を用いた再投影を行い，再投影誤差を比較することで距離を選択する．内部パラメータと外部パラメータを用いて再投影を行うことが出来るが，誤差の累積により再投影を正確に行うことが困難になる可能性がある．RAFT では視差の推定を行うため，誤差が累積することなく再投影を行うことが出来る．MiDaS では，視差を推定していない．そこで，スケール調整で得られた絶対距離から擬似的な視差を推定し，再投影を行う．再投影誤差が類似している場合，RAFT で得られた距離を選択する．再投影を行う際，テクスチャが弱い領域で行うため多くの場合，路面や空の領域に対して行う．実験的結果より，図 2に示すような電灯のような細い計測物が並んでいる場合，RAFT では図 3(a) に示すように空を誤計測してしまう．一方，MiDaS では空の領域を計測しない可能性が高い．そこで再投影を行う際，そのピクセルを MiDaS が計測していない場合，そのピクセルでは MiDaS の距離を選択する．また，どちらもしきい値以上の距離を計測している場合，計測値なしとする．

3.　精度評価実験

3.1　実験条件

本実験では，自動運転用のデータセットである DrivingStereo を用いて精度評価実験を行った [12]．DrivingStereo は，図 6に示すような様々な条件下で取得されたステレオ画像や真値の距離画像，キャリブレーションデータが含まれている．真値の距離と計測した距離の RMSE（Root Mean Squared Error）を比較した．図 6(b) における真値の距離画像を図 7に示す．近いほど赤く，遠いほど青いことを表している．DrivingStereo で取得された距離は約 80m まで計測されているため，80m までの距離を用いて評価した．また，真値の距離と計測した距離のどちらも計測が行えている場合のみ RMSE の計算に含めた．



(a) 提案手法

(b) RAFT

(c) MiDaS

図 8　エラーマップ

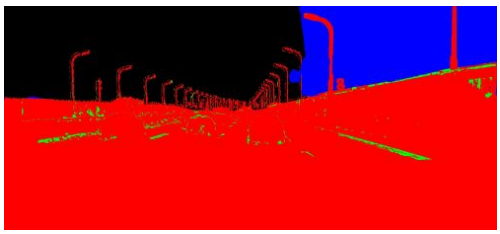


図 9 距離の選択（赤：RAFT，青：MiDaS，緑：ブロックマッチング，黒：計測していない）

3.2　実験結果

実験結果を図 8，9と表 2-5に示す．図 8は図 6(b) で実験を行った際のエラーマップを表しており，近いほど誤差が小さく，遠いほど誤差が大きいことを表している．また，真値の距離との差の絶対値を用いた．図 9は図 6(b) で実験を行った際，どの距離を選択したかを表している．赤色は RAFT，青色は MiDaS，緑色はブロックマッチングで得られた距離を選択したことを表している．また，黒色は計測していないことを表している．表 1は DrivingStereo 全体のテスト画像で実験を行った際の RMSE の平均とその標準偏差を表している．また，表 2-5は図 6に示す環境別に実験を行った際の RMSE の平均とその標準偏差を表している．図 8より，提案手法

表 1　全体における RMSE の平均と標準偏差

	平均 [m]	標準偏差 [m]
提案手法	3.91	1.08
RAFT	3.94	1.09
MiDaS	7.26	2.14

表 2　環境 1 における RMSE の平均と標準偏差

	平均 [m]	標準偏差 [m]
提案手法	3.77	0.64
RAFT	3.78	0.65
MiDaS	7.76	2.00

表 3　環境 2 における RMSE の平均と標準偏差

	平均 [m]	標準偏差 [m]
提案手法	4.31	1.29
RAFT	4.31	1.31
MiDaS	7.72	1.98

表 4　環境 3 における RMSE の平均と標準偏差

	平均 [m]	標準偏差 [m]
提案手法	4.38	0.82
RAFT	4.43	0.82
MiDaS	6.69	1.74

表 5　環境 4 における RMSE の平均と標準偏差

	平均 [m]	標準偏差 [m]
提案手法	2.88	0.86
RAFT	2.90	0.87
MiDaS	7.31	2.75

と RAFT では誤差が小さいことを表している赤色や黄色の領域が多いことが分かる．一方，MiDaS では誤差が大きいことを表している緑色や青色の領域が多いことが分かる．また，図 9より，提案手法では，多くのピクセルにおいて RAFT の距離を選択していることが分かる．また，空の領域において MiDaS の距離が選択されており，RAFT の誤計測を除去していることが分かる．

表 1より，提案手法における RMSE の平均と標準偏差が RAFT よりも低減したものの同等であることが分かる．また，MiDaS の RMSE の平均と標準偏差は提案手法や RAFT よりも低いことが分かる．表 2-5より，環境によらず計測を行うことが出来ていることが分かる．提案手法の距離精度が向上した理由として，ブロックマッチングで得られた距離を選択したことが原因として挙げられる．図 8(a)，(b) と図 9より，ブロックマッチングで得られた距離を選択している領域において，赤色になっており，誤差が低減したことが分かる．そのため，ブロックマッチングで得られた距離によって提案手法の距離精度が向上したと考えられる．

4.　結論

本論文では，ステレオビジョン手法である RAFT と単眼深度推定手法である MiDaS の融合を行う手法を提案した．実験によって，空のような領域の誤計測を行うことなく距離精度を向上させることが出来た．しかし，大幅な距離精度の向上を行うことが出来なかった．

今後の展望として，条件の変更による距離精度の変化の検証を行う．また，スケール調整のアルゴリズムの変更による距離精度の向上を行う．

参　考　文　献

[1] J. W. Perng, P. Y. Liu, K. Q. Zhong and Y. W. Hsu, “Front object recognition system for vehicles based on sensor fusion using stereo vision and laser range finder”, 2017 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW), pp. 261-262, 2017.

[2] Nobuo Sasaki, Naoyuki Iijima and Daiki Uchiyama, “Development of ranging method for inter-vehicle distance using visible light communication and image processing”, 2015 15th International Conference on Control, Automation and Systems (ICCAS), pp. 660-670, 2015.

[3] H. Hirschmuller, “Accurate and efficient stereo processing by semi-global matching and mutual information”, 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), pp. 807-814, 2005.

[4] Lipson, Lahav and Teed, Zachary and Deng, Jia, “RAFT-Stereo: Multilevel Recurrent Field Transforms for Stereo Matching”, International Conference on 3D Vision (3DV), 2021.

[5] Reiner Birkl and Diana Wofk and Matthias M¨uller, “MiDaS v3.1 –A Model Zoo for Robust Monocular Relative Depth Estimation”, arXiv preprint arXiv:2307.14460, 2023.

[6] Ibraheem Alhashim and Peter Wonka, “High Quality Monocular Depth Estimation via Transfer Learning”, arXiv e-prints arXiv:1812.11941, 2018.

[7] M. P. Muresan, M. Raul, S. Nedevschi and R. Danescu, “Stereo and Mono Depth Estimation Fusion for an Improved and Fault Tolerant 3D Reconstruction”, 2021 IEEE 17th International Conference on Intelligent Computer Communication and Processing (ICCP), pp. 233-240, 2021.

[8] Luo, Yue and Ren, Jimmy and Lin, Mude and Pang, Jiahao and Sun, Wenxiu and Li, Hongsheng and Lin, Liang, “Single View Stereo Matching”, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 155-163, 2018.

[9] Tosi, Fabio and Aleotti, Filippo and Poggi, Matteo and Mattoccia, Stefano, “Learning Monocular Depth Estimation Infusing Traditional Stereo Knowledge”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9799-9809, 2019.

[10] Moritz Menze and Andreas Geiger, “Object scene flow for autonomous vehicles”, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[11] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus, “Indoor segmentation and support inference from rgbd images”, In Computer Vision – ECCV 2012, pp. 746-760, 2012.

[12] Yang, Guorun and Song, Xiao and Huang, Chaoqin and Deng, Zhidong and Shi, Jianping and Zhou, Bolei, “DrivingStereo:A Large-Scale Dataset for Stereo Matching in Autonomous Driving Scenarios”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.