単眼カメラによる三次元環境認識*

Understanding 3D Semantic Scene with Monocular Camera

成岡 健一 Kenichi NARIOKA
西村 裕紀 Hiroki NISHIMURA
板持 貴之 Takayuki ITAMOCHI
猪俣 哲平 Teppei INOMATA

One of the essential features that autonomous driving systems and advanced driver assistant systems should have is an ability of understanding traffic scenes. In this paper, we propose a method to understand traffic scenes in detail with monocular camera, which can contribute to make sensing systems more useful and less expensive. We designed and trained Deep Neural Networks (DNNs) for semantic segmentation and monocular depth estimation to figure out semantic and geometric information of traffic scenes. Data for learning were gathered with a test vehicle with cameras that covers 360-degree and Velodyne LiDAR. Images were manually annotated using classes tailored for traffic scenes for semantic segmentation. Experimental result shows the trained network can accurately classify each pixel and also accurately estimate depth of the pixel of images in validation data. Global average of semantic segmentation reached 96.4%, while overall accuracy of depth estimation was 88.7%. We also developed a tool using a head mount display for virtual reality, that enable us to evaluate the result of estimation intuitively, helping us to check how well the estimation of proposed DNNs is.

Key words :
Safety, Active safety, Image processing, Monocular camera, Semantic segmentation, Depth estimation (C1)

1. まえがき

自動運転や高度運転支援システムには，走行環境を広く見渡し「何が」「どこに」あるのか認識する周辺センシング機能が不可欠である．車載システムがドライバの認知機能を代行あるいは補助することで，車の利便性や安全性の向上が期待される．

センシング機能を二次元的な意味認識（「何が」に相当）と三次元的な構造認識（「どこに」に相当）に分けたとき，前者に有用なデバイスがカメラである．物体検出など画像認識機能を備えたセンシングカメラの技術発展が進み，市販車への搭載が急速に増加している[1)]．他方，後者の三次元認識にはステレオカメラ[2)]，ミリ波レーダ，レーザレーダなどが主に使用される．これらのデバイスは機能とコストのトレードオフがあり，高いレベルでのバランスが要求される．すなわち，走行環境中の多種多様な物体を，詳細に，また広範囲に渡って認識することが求められる一方で，システムのメンテナンス性の悪化やコストの増大を招くため，やみくもにデバイスの数や種類を増加させるのは好ましくない．

本論文では，低コストで高機能なセンシングを実現するための要素技術として，深層ニューラルネットワーク（Deep Neural Networks, 以下 DNNs）を基盤とした，単眼カメラによる環境認識手法を提案する．Fig. 1 に提案手法の概略を示す．単眼カメラ画像を入

＊（公社）自動車技術会の了解を得て「2017 年秋季大会学術講演会 講演予稿集 No.117-17 P.129 ～文献番号 20177623」より一部加筆して転載

力とし，二次元の意味認識を担うセマンティックセグメンテーションと三次元の構造認識を担う単眼デプス推定の結果を統合してシーン認識を行う．セマンティックセグメンテーションは多クラスの分類をピクセル単位で詳細に行い，デプス推定 DNN は奥行きをピクセル単位で詳細に推定する．さらに，認識の広範囲化のため，全方位の教師データを用いてそれらの DNN を学習させる．

関連研究として，Schneider et al.[3] はステレオカメラによるセマンティックセグメンテーションとデプス推定の結果をスティクセル表現により統合している．また，Song et al.[4] は画像とデプスマップを入力とし，三次元ボクセルに対するセグメンテーションを出力する手法を提案している．これらに対し，本論文では単眼カメラでのデプス推定を行い，かつ認識を全方位方向に拡張するという点に新規性がある．

2 節ではセマンティックセグメンテーション，3 節ではデプス推定のタスクと実験結果について述べ，4 節では VR ディスプレイを用いた独自の評価ツールについて紹介する．

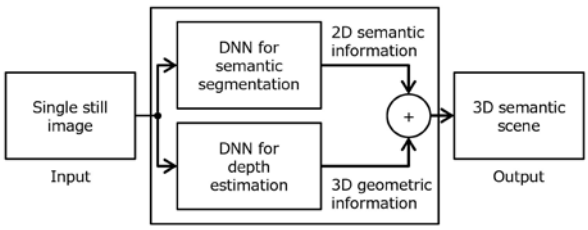

Fig. 1 Overview of DNNs for understanding 3D semantic scene. 2D semantic information and 3D geometric information are inferred by DNNs and integrated to get 3D semantic structure of environment

2. セマンティックセグメンテーション

2.1 タスク概要

セマンティックセグメンテーションはピクセルラベリングとも呼ばれ，画像中の各ピクセルに対し，事前に定義したクラスの中からひとつを選択して割り当てるタスクである．深層学習技術の発展にともない画像認識分野で盛んに取り組まれており，CamVid [5] や Cityscapes [6] など車載画像の公開データセットを用いたベンチマークも活発化している．

2.2 DNN 構造

セマンティックセグメンテーションの手法として畳み込みニューラルネットワークを用いた教師あり学習が主流であり，中でも SegNet [7] に代表されるエンコーダーデコーダ型が注目されている．前半のエンコーダ部で畳み込み演算とプーリング処理を繰り返して情報を圧縮し，後半のデコーダ部ではアップサンプリングと畳み込み演算を繰り返しながら段階的に解像度を高め，ソフトマックス層を介して各ピクセルの推定クラスを出力する．

一般に，認識精度は計算負荷とトレードオフの関係にある．限られた車載計算資源を有効に利用するため，層数や各層のマップ数を調整してパラメタ数を削減し，精度と処理時間の両立を狙う DNN を設計した．その構造を Fig. 2 に示す．

2.3 学習データ

全方位カメラ（PointGrey 社製 Ladybug5）をルーフ上に設置した実験車両を用いて，テストコース内で画像データを収集した．Fig. 3 に示すように，円周方向に等間隔に設置されたカメラ 5 台により全方位がカバーされている．歩行者，車両，道路など，走路環境に特化したクラス分類を定義し，収集した画像に対してアノテーションを行った．

2.4 結果

アノテーション付き画像データを訓練データと検証データに分割し，教師あり学習により DNN を訓練させた．結果の一例を Fig. 4 示す．2 つのシーンについて，それぞれ上段から順に入力画像，真値ラベルマップ，推定ラベルマップが示されている．各クラスは予め割り当てられた色で表現されている．尚，推定はカメラ毎，フレーム毎に独立に行っており，左の列から順に，左後方，左前方，正面，右前方，右後方に取り付けられたカメラ画像入力に対する結果を示している．

結果より，全周に渡って車両，歩行者，道路などが詳細かつ高精度に認識されている様子が確認できる．また，道路領域と非走行領域が適切に区別できている点は注目に値する．これは畳み込み演算によって周辺のコンテキストが適切に参照されている効果と考えら

れる．検証データに対する定量的な評価結果は，ピクセル全体精度 96.4%，平均クラス精度 88.7%，平均 IoU は 65.4% となった．

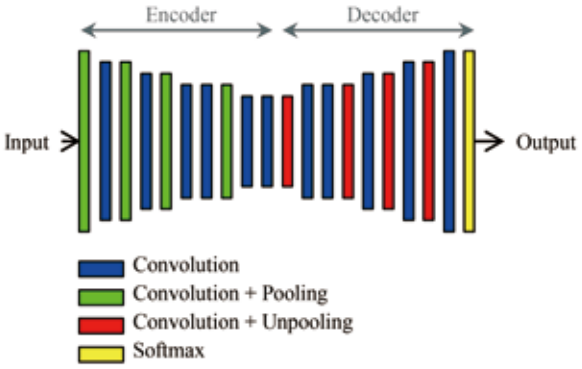

Fig. 2 Proposed DNN for semantic segmentation. The encoder part has iterative convolution al layers and pooling layers, while the decoder part has iterative convolutional layers and unpooling layers

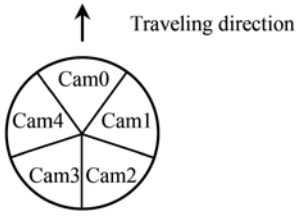

Fig. 3 Configuration of cameras. All cameras are placed equiangularly. Cam0 faces the front of the test vehicle

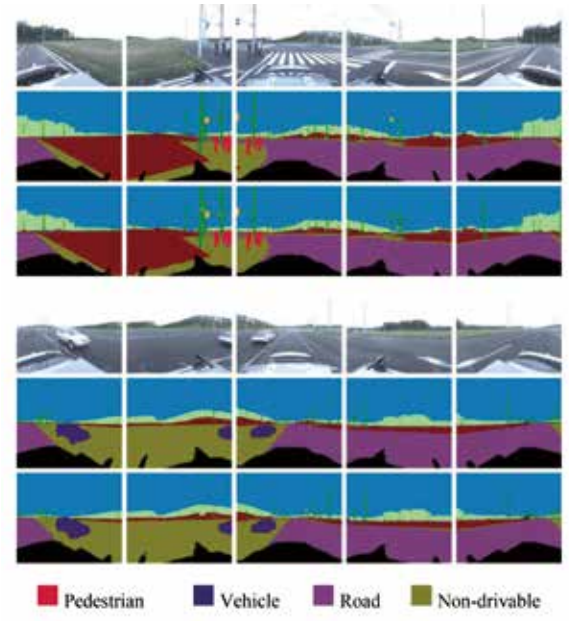

Fig. 4 Result of semantic segmentation. Input images, manually annotated label maps (ground truth), and estimated label maps are shown in the top, middle, and bottom, respectively in each figure. Each class is colorized with its pre-defined color, which is shown in the bottom

3. 単眼デプス推定

3.1 タスク概要

デプス推定は，画像から各ピクセルの奥行きを推定するタスクである．ステレオ視はこの代表的な手法のひとつで，複数のカメラでとらえた特徴点をマッチングし，カメラ間の位置関係から奥行きを求める．また，単眼カメラの画像列とカメラモーションを利用する Structure from Motion(SfM) 法もよく知られた方法である．これらの三角測量の原理に基づく手法に対し，教師あり学習によって画像と奥行きの対応関係を直接学習する手法が提案されている．複数のカメラを必要としないためハードのコストが抑えられ，カメラ間の調整や経時変化への対応も不要であるという利点がある．また，自車が停止しているケースや，周辺に移動物が多く存在するケースなど SfM の苦手とする状況でもデプス推定が可能である．

3.2 DNN 構造

Eigen et al.[8] により提案されたマルチスケール畳み込みニューラルネットワークは，単眼デプス推定の有力手法のひとつである．全体が 3 つのスケールで構成され，各スケールには前段スケールの出力と元画像が入力される．スケール 1 は画像特徴量を，スケール 2 では推定デプスマップを，スケール 3 では高解像度の推定デプスマップをそれぞれ出力する．

車載計算資源の制限から，認識精度を極力維持しつつ計算量やサイズを低減することが重要である．本論文では，層数や各層のマップ数の調整，Firemodule[9] で畳込み層を置き換えるなどの改善を行い，精度と処理時間の両立を狙った．設計した DNN の構成を Fig. 5 に示す．

3.3 学習データ

全方位カメラと全方位 LiDAR（Velodyne 社製）をルーフに取り付けた実験車両を用いて，テストコース内で画像および点群データを収集した．得られた点群を各カメラの画像座標系に投影し，画像に対応する真値デプスマップを生成した．

3.4 結果

画像と真値デプスマップのペアを訓練データと検証データに分割し，教師あり学習により DNN を訓練させた．結果の一例を Fig. 4 示す．2 つのシーンについて，それぞれ上段から順に入力画像，真値デプスマップ，推定デプスマップを示す．奥行きはヒートマップで表され，近距離が赤色，遠距離が青色で表現されている．尚，推定はカメラ毎，フレーム毎に独立に行っており，左の列から順に，左後方，左前方，正面，右前方，右後方の各カメラ画像に対する結果を示している．

Fig. 6 より，LiDAR 真値に近いデプスを良好に推定できていることが確認できる．次に，定量評価結果を Table 1 に示す．評価法は先行研究[4]に準拠し，画像中のピクセル i のデプス真値を d_i^*，デプス推定値を d_i するとき，$\max(d_i^*/d_i, d_i/d_i^*)<1.25$ となるピクセルを正解として平均正解率を算出した．全体平均は 88.7% で，各カメラに偏ることなく，全方位に渡り高い精度でデプスが推定されていることが示された．

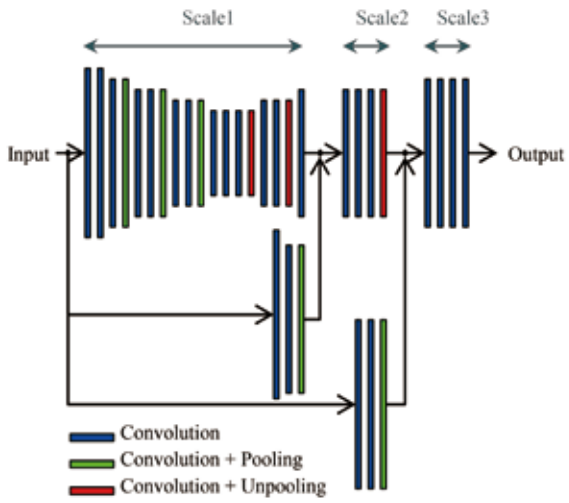

Fig. 5 Proposed DNN for monocular depth estimation. Scale1 extracts feature of an input image. Scale2 takes the output of the Scale1 together with the input filtered by several convolutional and pooling layers. Scale3 has a function of upsampling to make the resolution of the output higher

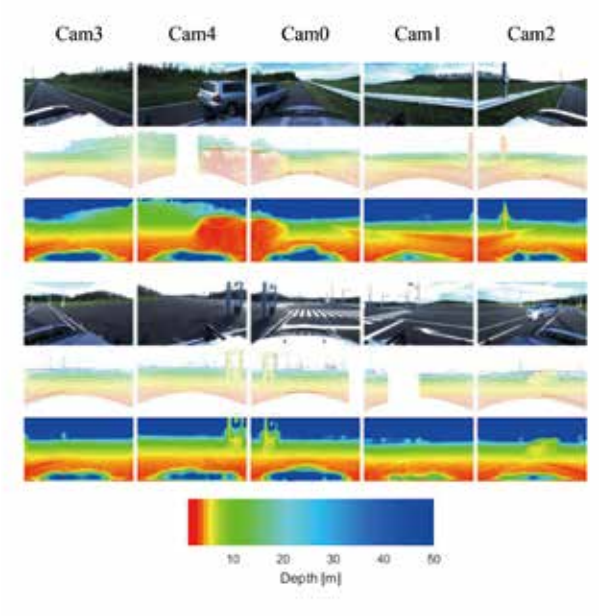

Fig. 6 Result of depth estimation. Input images, depth maps measured by Lidar, depth maps estimated by the proposed network are shown in the top, middle, and bottom, respectively in each figure

Table 1 Accuracy of depth estimation

Camera	Cam3	Cam4	Cam0	Cam1	Cam2
Accuracy[%]	87.2	87.8	88.4	87.0	88.3

4. 認識評価ツール

セマンティックセグメンテーションやデプス推定の結果はラベルマップ（Fig. 4）やヒートマップ（Fig. 6）のように画像として平面的に表示させる手法や，三次元点群表示ソフト等を利用して立体的に表示させる方法が一般的である．本論文では，推定結果をより直感的に評価するため，VR ヘッドマウントディスプレイ（以下 VR ディスプレイ）を利用した三次元表示ツールを開発した．

入力画像サンプルと三次元表示ツールの映像例を Fig. 7 に示す．ただし，実際に VR ディスプレイを装着すると，両眼視差を利用した遠近感が加わる．Fig. 7 (a) は 5 台のカメラで撮像された画像を示す．Fig. 7 (b) はそれらの画像を，各カメラ位置を考慮して VR 空間内に並べたものである．Fig. 7 (c) はセマンティックセグメンテーションの推定結果を重畳表示している．Fig. 7 (d) では，デプス推定結果とカメラパラメータから各ピ

クセルの三次元位置を求め，VR 空間に表示させている． 各点は，VR ディスプレイ装着者の頭部の動きに合わせて移動する．Fig. 7 (e) および Fig. 7 (f) は視点の移動による見え方の変化の様子を示している．

全周カメラの認識結果を統合することで切れ目のない三次元世界が構成され，VR の生じる遠近感の効果から，装着者はその世界に入り込んだような感覚を得る．このツールを通して推定結果を見ることで，物体形状の自然さ，物体間境界の正確さなどが直観的に確認できる．また，セマンティックセグメンテーションとデプス推定の統合を検討する上でも，本ツールは有用である．一例として，物体領域毎にデプス値を補正する処理を入れることで推定精度を改善できることが確認されている． また，デプス情報を利用したセマンティックセグメンテーションの改善事例も報告されているように[10]，様々な手法で認識結果の相補的な改善が期待される．

5. まとめ

本論文では，セマンティックセグメンテーションとデプス推定を組み合わせた，単眼カメラによる環境認識手法を提案した．両タスクともに，全周データを用いて DNN を学習した結果，全方位に渡り良好な推定結果が得られた．車載カメラに関する公開データベースでは前方カメラ画像のみを対象としたものがほとんどで，全周カメラ画像での学習と評価に拡張した点が本論文の貢献のひとつである．また，VR ディスプレイを利用した三次元表示ツールを開発し，三次元環境認識の新しい評価形態を提示した．

単眼カメラによるセンシングの高度化は，カメラ単独でシステムを構成する際に有用であることはもちろんであるが，他のセンシングデバイスとのセンサフュージョンの品質向上にも効果が高い．低コストで高機能なセンシング技術の開発を通じ，高級車だけでなく大衆車にも安全システムの裾野を広げ，安全な車社会の実現に貢献することを目指す．

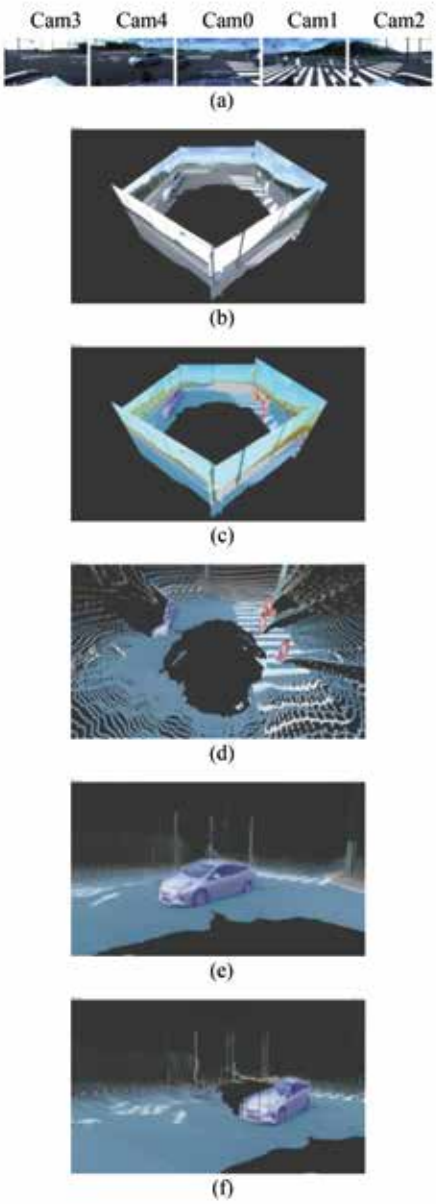

Fig. 7 3D evaluation tool. (a) and (b) show input images from five cameras. (c) Result of semantic segmentation. (d) Result of semantic segmentation and depth estimation. (e) and (f) show the same vehicle from the different points of view

参考文献

1) 二反田直已，公文宏，玉津幸政：車載カメラによるセンシング・認識技術，自動車技術，自動車技術会，Vol. 66, p. 67-72, 2012
2) 実吉啓二：ステレオカメラによる自動車運転支援システム, CVIM, 2013
3) Lukas Schneider, Marius Cordts, Timo Rehfeld, David Pfeiffer, Markus Enzweiler, Uwe Franke, Marc Pollefeys, Stefan Roth: Semantic Stixels: Depth is Not Enough, IEEE Intelligent Vehicles Symposium, 2016
4) Jonas Uhrig, Marius Cordts, Uwe Franke, Thomas Brox: Pixel-level encoding and depth layering for instance-level semantic labeling, German Conference on Pattern Recognition, 14-25, 2016
5) Gabriel J Brostow, Julien Fauqueur and Roberto Cipolla Semantic object classes in video: A high-definition ground truth database Pattern Recognition Letters, Elsevier, Vol. 30, 88-97 2009
6) Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth and Bernt Schiele The Cityscapes Dataset for Semantic Urban Scene Understanding Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016
7) Vijay Badrinarayanan, Alex Cendall, Roberto Cipolla:SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017
8) David Eigen, Rob Fergus: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture, Proceedings of the IEEE International Conference on Computer Vision, 2015, 2650-2658
9) Forrest N Iandola, Matthew W Moskewicz, Khalid Ashraf, Song Han, William J Dally and Kurt Keutzer SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and 1MB model size arXiv preprint arXiv:1602.07360, 2016
10) Jonas Uhrig, Marius Cordts, Uwe Franke, Thomas Brox: Pixel-level encoding and depth layering for instance-level semantic labeling, German Conference on Pattern Recognition, 14-25, 2016

著者

成岡 健一
なりおか けんいち

ADADAS 技術１部　博士（工学）
画像認識の先行技術開発に従事

西村 裕紀
にしむら ひろき

ADADAS 技術１部
画像認識の先行技術開発に従事

板持 貴之
いたもち たかゆき

ADADAS 事業部
画像認識の先行技術開発に従事

猪俣 哲平
いのまた てっぺい

株式会社モルフォ　博士（工学）
深層学習を用いた画像認識技術の研究開発に従事