自動運転向けディープラーニング障害物検出

Multiple Obstacle Detection for Autonomous Car

築澤宗太郎 Sotaro Tsukizawa
グレゴリー セネー Gregory Senay
ミンヤン キム Min-young Kim
ルカ リガッツィオ Luca Rigazio

要　旨

自動運転車の障害物検出機能のための新たな物体検出技術Temporal Faster R-CNN（Region-based Convolution Neural Networks）を提案する．一般的な複数物体検出技術は静止画を対象としているため，動画中の物体検出であってもフレームごとに個別に物体を検出するが，自動運転向けの障害物検出では車載カメラからの動画が認識対象であるため，認識性能に課題があった．本技術は時系列情報，すなわち過去のフレームの情報と現在のフレームの情報を同時に認識に用いることで12.9 %の演算量増加で7.7 %～17.9 %の認識率向上を確認した．

Abstract

This paper presents the novel approach of a Multiple Object Detection (MOD) system for an autonomous car. Our development is based on time-series data and is an extension of a state-of-the-art MOD system called “Faster R-CNN.” The experiments focused on the KITTI dataset, which is recorded in real-driving conditions in a city, rural areas and highways. In our approach, the model is able to take into account temporal information that is provided by previous frames. The results show an improvement in performance while the computational cost is kept almost the same.

1. はじめに

近年，自動運転車の実用化，実証実験が盛んになっている．アメリカではすでに公道を実証実験車両が走行しており，日本でも，2020年までに限定区間，2030年までに全区間での完全自動運転を実現するよう官民を挙げて取り組んでいる．

自動運転における重要な機能として，路上の障害物を検出する障害物検出機能がある．すでに物体の位置を検出する技術はLIDAR（Light Detection and Ranging）センサ，ステレオカメラなど距離センサによって実用化されているが，これらは物体の位置は検出できるものの，物体の種類を識別できないという課題がある．そのため，例えば，横断歩道脇にある物体が看板なのか歩行者なのかを区別できず，自動運転で安全に走行するためにはたとえ看板であっても緊急停止できるように毎回減速しなくてはならないという問題がある．このような問題を解決する手段として，近年，車載カメラを用いたディープラーニング複数物体検出技術が注目されている．

複数物体検出技術は，画像中の物体の位置と種類を認識する画像認識技術の1つであり，距離センサでは困難な，物体の種類を認識できる技術として有望視されている．ディープラーニングは機械学習の一手法で，近年圧倒的な認識性能で注目されている技術であり，音声認識，画像認識，自然言語処理などですでに多くの実績を上げている．複数物体検出技術においても高い認識性能を実現している．しかし，ディープラーニングには高い認識性能と引き換えに，膨大な演算量が必要であり，車載LSIで動作困難であるという問題がある．この問題の解決のため，筆者らはディープラーニングの高い認識性能と低演算量を両立する自動運転向け画像認識技術を開発している．

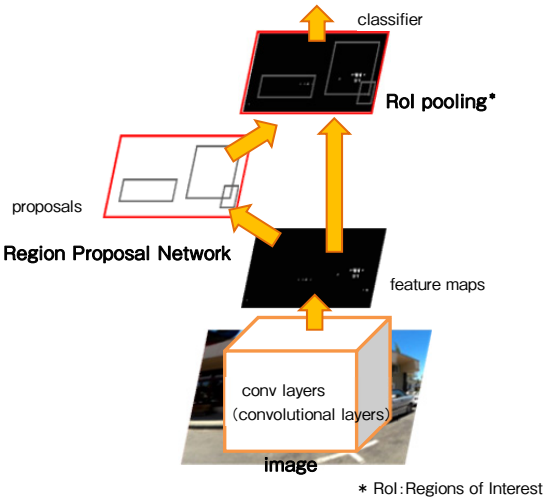
本稿では，自動運転向け複数物体検出技術Temporal Faster R-CNNについて述べる．

2. 複数物体検出技術

2.1 従来のディープラーニング複数物体検出技術

ディープラーニングを用いた複数物体検出の代表的な技術としてR-CNN（Region-based Convolution Neural Networks）[1]，Fast R-CNN[2]がある．これらは1) 物体候補領域の検出，2) 各候補領域に対するクラス識別 の2段階処理によって検出処理を行う．しかし，1)の処理において，Selective Searchなどのアルゴリズムを用いているため物体候補領域の検出の演算量が非常に大きい．そのため，PC上で実行しても1枚の画像当たりの検出処理に約2秒かかってしまい，リアルタイム検出が求められる車載カメラへの適用は困難である．この物体候補領域検出の処理速度問題を解決する手法として，Faster R-CNN[3]がある．これは，**第1図**に示すように，Faster R-CNNにおける画像のfeature map化と物体候補領域の検出をCNNで同時に行うRegion Proposal Networkによって実現する手法で，R-CNNに対して認識率を向上したうえで検出速度を大幅に高速化している．**第1表**にR-CNNとFaster R-CNNの物体検出技術評価データセットPASCAL-VOC（Pattern Analysis, Statistical Modelling and Computational Learning-

Visual Object Classes）での処理速度と認識率を示す．検出速度がR-CNNの0.5 fpsに対してFaster R-CNNは5 fps，さらに認識率（MAP：Mean Average Precision）が66.9 %から69.9 %に向上している．しかし，Faster R-CNNの5 fpsでも例えば時速60 kmで走行する場合に，障害物が出現してから発見するまでに約3.3 m進んでしまい，障害物の衝突回避には不十分である．このように従来技術ではディープラーニングを用いたリアルタイム障害物検出は実現が困難であった．



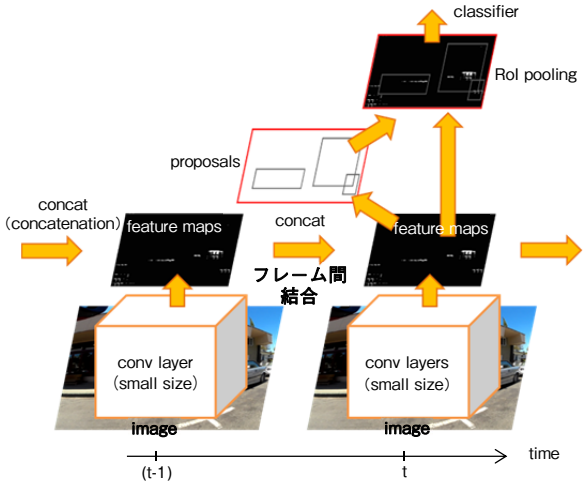
第1図　Faster R-CNNアーキテクチャ
Fig. 1　Faster R-CNN architecture

第1表　R-CNNとFaster R-CNNの速度比較
Table 1　Comparison between R-CNN and Faster R-CNN
Dataset：PASCAL-VOC 07,
GPU：NVIDIA GeForce GTX1080 (注1)

Method	R-CNN	Faster R-CNN
speed	0.5 fps	5 fps
MAP	66.9 %	69.9 %

2.2　Temporal Faster R-CNN

Faster R-CNNをさらに高速化させ，車載カメラでの障害物検出を実現するため，筆者らは，車載カメラのデータが，一般的に物体検出技術が対象としている静止画ではなく，動画であることに着目し，Faster R-CNNを動画に対応せることでさらに高速化させるTemporal Faster R-CNNを開発した． 2.1節で述べたように，複数物体検出において，演算量の観点で最も課題となるのは，feature map算出である．feature mapには，物体の候補位置と物体の種類に関する情報が含まれており，認識性能に大きな影響を与える．そのため，これを軽量化すると認識性能が大きく低下するため実用性に課題があった．本手法では，feature map算出を軽量化し，さらにマルチフレームを用いる，すなわち，過去のフレームの検出で算出したfeature mapを新たなフレームの検出にも用いることでこの問題の解決を図る．一般的に認識率と演算量はトレードオフの関係にあるが，本手法では，feature map算出にSqueezeNet[4]を適用することで演算量削減を行い，これによって悪化した認識性能をマルチフレームのfeature map利用により向上させる．これより，feature map算出処理を軽量化した場合でも認識率の低下を抑制することができ，結果的に同等の認識率で高速化を図ることができる．また，車載カメラでの障害物検出においては，駐車車両で体の大部分が隠された歩行者のような物体の部分隠蔽や，トンネル出入り口などでの明るさの変化により静止画だけでは認識が困難な状況が頻発する．これらに対し，マルチフレームfeature mapを用いることでシングルフレームの静止画だけでは検出困難な状況でも動作することが期待できる．**第2図**にTemporal Faster R-CNNのアーキテクチャを示す．



第2図　Temporal Faster R-CNNアーキテクチャ
Fig. 2　Temporal Faster R-CNN architecture

第2図に示すように，本手法では，過去に検出のfeature mapを保持し，新たなフレームのfeature mapと結合してRegion Proposal NetworkとROI（Regions of Interest）poolingに入力する．このとき，使用する過去フレームは1フレームに限定されず，複数フレームのfeature mapを使用することができる．また，使用する過去フレームは連続したフレームである必要はなく，検出したい物体や車載カメラ自身の移動速度などに応じてフレーム間隔を変更することができる．例えば，歩行者のみを検出したい場合，歩行者の移動速度はそれほど速くないため，カメラの一

（注1）GeForce GTXはNVIDIA Corp.の登録商標．

般的なフレームレートである30 fpsで撮影すると，ほぼ同じ画像になってしまい，過去フレームの使用によって増える情報量が少ない．そこで，例えば，3フレーム間隔で過去4フレーム分のfeature mapを使用し400 msに相当する時間軸の情報を使用することで，動きが遅い歩行者のような物体の検出でも効果を発揮することができる．

2.3 マルチフレームfeature map算出

第2図のconv layersで行う各フレームのfeature mapの算出は，VGG（Visual Geometry Group）[5]のような画像識別に使用されるネットワークから最終の識別層を除いたCNN部分を使用する．本手法では，これにSqueezeNetを適用する．SqueezeNetは，メモリーサイズが従来の1/50となる非常に小さいネットワークとして近年注目されているネットワークであるが，認識性能がVGGなどと比較して低いという問題がある．このため，各フレームで使用するconv layersはパラメータをシェアした同一のネットワークを用い，各フレームのfeature mapを結合することによって認識性能を向上させている．

本手法の特徴として，feature mapの算出において学習時と検出時で，動作が異なることが挙げられる．すなわち，学習時は，使用するすべてのフレームのデータを都度conv layersでfeature mapの算出を行うが，検出時は，新たなフレームのfeature mapのみconv layersで算出する．これにより，学習時の演算量は使用する過去フレーム数に応じて大きく増大するが，検出の演算量はほとんど変わらない．例えば，これは，ベースとなったアルゴリズムFaster R-CNNにおいて，feature map算出にVGGで行い，過去3フレーム分使用した場合，学習時間は現在のフレームだけで学習した場合と比較して約2.2倍に増大するが，検出時のフレームレートは5.6 %の増大にとどまる．これはfeature map算出の演算量が大半を占めていることに起因する．すなわち，feature mapの演算量は全体の約3/4を占めており，feature map結合のための処理量は1/10以下であるため，この処理を追加することによる検出時の総演算量の増加は10 %以下にとどまるためである．

3. 実験

本提案手法の有効性を示すため，車載カメラ映像を用いたデータセットKITTI MOD（Multiple Object Detection）dataset[6]による評価実験を行う．本実験では，マルチフレーム使用時の認識性能向上効果を示すため，feature map算出にVGGとSqueezeNetを用いた場合について，それぞれ単フレームのみを使用した場合とマルチフレームを使用した場合の比較を行う．

3.1 KITTI MOD dataset

The KITTI MOD datasetは，実写映像による複数物体検出ベンチマークデータセットである．映像は，ドイツの中規模都市周辺で撮影され，都市部，郊外，高速道路の映像が含まれている．映像中には最大15台の車両と30人の歩行者が含まれており，7481枚の学習用画像と7518枚のテスト用画像で構成されている．また全データセット中には80256個のラベル付き物体が含まれている．

本実験では，同データを使用しているYu Xiangら[7]，他[8]-[10]と同様，学習用画像として提供されているデータを分割した3682枚のミニ学習用画像，3799枚のミニテスト画像を用いて実験を行う．これは，7518枚のテスト用画像を用いた評価実験はKITTIのオンラインサイトで行われるが，結果の提出が4週に1回しかできず，複数条件での実験が困難なためである．実験は，路上の障害物として検出優先度が高い車，歩行者，サイクリストの3種類の物体について行う．

3.2 実験条件

実験は，以下のパラメータにて行う．

- Learning rate : 0.005
- Learning rate step: 30000 iterations以降, 0.1倍
- Training input scales: [400, 600, 800, 1000]
- Maximum width: 2000
- Testing scale: 600
- Iteration: 90000

また，マルチフレーム使用条件の実験は，すべて現在のフレームと過去3フレームを使用する．

feature map算出に使用するネットワークとして，本実験ではVGG-16とSqueezeNetを使用する．VGG-16は，Faster R-CNNで用いられているネットワークで，本実験の比較対象として用いる．本実験では，マルチフレームのSqueezeNetをfeature map算出に用い，演算量削減と高認識率維持の両立が可能かを評価する．

3.3 実験結果

第2表から**第5表**に実験結果を示す．なお，Mono FrameとMulti Framesはそれぞれ，単フレーム，マルチフレームを指し，SQNはSqueezeNet, VGGはVGG-16を指す．また，表中のEasy，Moderate，Hardは，KITTI MOD datasetにあらかじめ付与されている検出難易度を示すラベルである．

第2表　KITTI MODミニテストデータにおける車クラスの認識率（MAP）
Table 2　MAP on KITTI MOD mini-val of car class

車	Easy	Moderate	Hard
Mono Frame SQN	80.47 %	68.63 %	56.43 %
Multi Frames SQN	**83.23 %**	**73.31 %**	**60.06 %**
Mono Frame VGG	**93.96 %**	**78.67 %**	67.94 %
Multi Frames VGG	91.93 %	78.32 %	**68.23 %**

第3表　KITTI MODミニテストデータにおける歩行者クラスの認識率（MAP）
Table 3　MAP on KITTI MOD mini-val on pedestrian class

歩行者	Easy	Moderate	Hard
Mono Frame SQN	55.07 %	45.53 %	42.75 %
Multi Frames SQN	**77.41 %**	**63.46 %**	**58.32 %**
Mono Frame VGG	**88.09 %**	**70.30 %**	**61.99 %**
Multi Frames VGG	86.43 %	68.86 %	61.58 %

第4表 KITTI MODミニテストデータにおけるサイクリストクラスの認識率（MAP）
Table 4　MAP on KITTI MOD mini-val on cyclist class

サイクリスト	Easy	Moderate	Hard
Mono Frame SQN	40.95 %	34.90 %	32.87 %
Multi Frames SQN	**51.59 %**	**43.29 %**	**40.54 %**
Mono Frame VGG	61.51 %	49.67 %	46.18 %
Multi Frames VGG	**63.22 %**	**52.95 %**	**50.15 %**

第5表　単フレームとマルチフレーム時の処理速度
Table 5　Speed using mono or multi-frames
（GPU： NVIDIA GeForce GTX1080）

処理速度	VGG	SQN
Mono Frame	3.90 fps	10.75 fps
Multi Frames	3.69 fps	9.52 fps

3.4　考察

実験結果に示すように，SqueezeNetでマルチフレームを用いた場合，すべての条件下で単フレームを用いる場合よりも認識性能が向上することを確認した（**第2表**～**第4表**）．例えば，検出難易度がModerateの場合，歩行者クラスでは+18 %，サイクリストクラスでは+9 %と大きく性能改善していることが確認できる（**第4表**）．なお，車クラスにおいて+3.86 %と性能改善が他クラスよりも低いことが確認できるが，これは，歩行者やサイクリストと比較して車の移動速度が速いため，10 fpsの本データセットではフレーム間の移動量が大きすぎたためと考えられる（**第2表**）．

VGG-16でマルチフレームを使用した場合，車クラスのHardや，サイクリストクラスで認識率の向上を確認できたが，その他のクラスについては若干の低下が発生した（**第2表，第4表**）．これは，VGG-16が巨大なネットワークであり，今回使用したミニ学習用画像のデータ規模では十分な学習が行えなかったためと考えられる．

なお，メモリー消費量は，現在フレームのみを使用した場合に比べて6 %の増加にとどまっており，メモリーサイズに関して実用上大きな影響はなかった．

以上の実験結果から，本提案手法によって，Faster R-CNN（本実験におけるVGG-16単フレーム）と比較して認識率低下を3.7 %～10.7 %に抑制したうえで約2.4倍の高速化を実現できる．また，マルチフレーム利用単体の効果として，単純な高速化手法であるFaster R-CNNのVGG-16部分をSqueezeNetに置き換えた場合と比較すると，マルチフレームを用いることで，12.9 %の演算量増加で7.7 %～17.9 %の認識率向上が可能であることを確認した．

4. まとめ

本稿では，自動運転向け複数物体検出技術として，動画に特化したアルゴリズムTemporal Faster R-CNNを提案した．本技術は，静止画を対象とした複数物体検出技術Faster R-CNNを拡張し，feature map算出にSqueezeNetを適用し，さらに時系列情報，すなわち過去の複数フレームを用いている．実験では，車載カメラを用いた実環境映像データKITTI MOD datasetを用い，路上障害物として認識優先度が高い車，歩行者，サイクリストの検出を行った．本提案手法の実験により，従来手法と比較して認識率を維持したうえで約2.4倍の高速化実現を確認した．

今後，時系列情報を用いた拡張をさらに行い，性能改善を行っていく．

参考文献

[1] Ross Girshick et al., “Region-based convolutional networks for accurate object detection and segmentation,” IEEE transactions on pattern analysis and machine intelligence 2016, vol.38, Issue.1, pp.142-158, 2016.
[2] R. Girshick, “Fast R-CNN,” International Conference on Computer Vision (ICCV), Santiago, Dec. 2015
[3] Shaoqing Ren et al., “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” NIPS2015 Proceedings of the 28th International Conference on Neural Information Processing Systems, pp.91-99, Dec. 2015.
[4] Forrest N. Iandola et al., “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 1MB model size,” arXiv preprint, arXiv: 1602.07360, Feb. 2016.
[5] K.Simonyan et al., “Very deep convolutional networks for largescale image recognition,” arXiv preprint arXiv:1409.1556, Sep. 2014.
[6] A.Geiger et al., “Are we ready for autonomous driving? The

KITTI vision benchmark suite," Computer Vision and Pattern Recognition (CVPR), Providence, Rhode Island, June 2012.

[7] Yu Xiang et al., "Data-driven 3d voxel patterns for object category recognition," Computer Vision and Pattern Recognition (CVPR), Boston, June 2015.

[8] Fan Yang, et al., "Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers," Computer Vision and Pattern Recognition (CVPR), Las Vegas, June 2016.

[9] Khalid Ashraf, et al., "Shallow Networks for High-Accuracy Road Object-Detection," arXiv preprint arXiv: 1606.01561, June 2016.

[10] Shuai Zheng, et al., "Conditional random fields as recurrent neural networks, " International Conference on Computer Vision (ICCV), Santiago, Dec. 2015.

執筆者紹介

築澤　宗太郎　Sotaro Tsukizawa
ビジネスイノベーション本部
AIソリューションセンター
AI Solutions Center,
Business Innovation Div.

グレゴリー セネー　Gregory Senay
パナソニックR&Dカンパニー アメリカ
パナソニック シリコンバレー研究所
Panasonic Silicon Valley Lab.,
Panasonic R&D Company of America
Computer science Ph. D.

ミンヤン キム　Min-young Kim
パナソニックR&Dカンパニー アメリカ
パナソニック シリコンバレー研究所
Panasonic Silicon Valley Lab.,
Panasonic R&D Company of America

ルカ リガッツィオ　Luca Rigazio
パナソニックR&Dカンパニー アメリカ
パナソニック シリコンバレー研究所
Panasonic Silicon Valley Lab.,
Panasonic R&D Company of America