「AVOD に基づいたマルチスケール特徴量を用いたオブジェクト検出に関する研究」
A study on object detection using multi-scale features based on AVOD

経営システム工学専攻 王　勇翔
Management Systems Engineering Yongxiang Wang

1　はじめに

オブジェクト検出は 2012 年から Deep Learning 時代に入り、勾配ベース手法から Convolutional Neural Network (CNN) を利用した手法に大きく移っている。オブジェクト検出が 2D 画像に大きな進歩があるけど、自動運転や拡張現実（AR、augmented reality）などの多くのアプリケーションで 3D の理解に強く求められている。最近、3D センサーがモバイルデバイスや自動運転の車に導入され、ますます多くの 3D データが収集し、処理される。代表的な 3D データでは、Lidar というセンサーから収集した点群データである。RGB 画像と比較すると、点群には独自の特性がある。一方では、相対的な位置と正確な深さの構造的および空間的情報を提供している。一方、それぞれは無秩序、疎ら、また局所に敏感である。本研究は MV3D[3] が提案した BEV マップを Lidar データの特徴にする。オブジェクトの物理的なサイズは、BEV に投影したときに保持される。また、BEV のオブジェクトは異なるスペースを占有するため、オーバーラップに関する問題なくなどメリットがある。

最近では、車など大きなオブジェクトが表現が良いの手法が提案されている。しかし、歩行者など小さなオブジェクト検出の表現は大きなオブジェクト比べるとまだ差がある。本稿は AVOD をベースに、小サイズオブジェクト検出に向け新たなオブジェクト検出モデルを提案した。

2　多尺度特徴融合検出モデル

融合検出モデルの構造を図 1 に示す. まず、画像データを入力し、特徴生成ネットワークから三つ尺度の画像特徴マップを生成する。Lidar データから BEV マップを生成し、特徴生成ネットワークに入力する。三つ尺度の BEV 特徴マップを得る。両方の特徴マップ解像度が一番高い特徴マップを切り取り、融合する。[3] と同じ区域提案ネットワーク（RPN（region proposal network））に入力し、3D 提案アンカーを生成する。画像と BEV 両方三つ尺度の特徴マップそれぞれ切り取り、融合する。三つの全結合層を構築し、3D 提案アンカーを用いて、融合した 3 つの特長マップからオブジェクトのバウンディングボックスと信頼度を得る。オブジェクトを検出する。

2.1　特徴生成ネットワーク

特徴生成ネットワークの前半部分は darknet-53 をベースに構築した。特徴生成ネットワークの後半部分は特徴ピラミッドを用いた構築した。提案融合モデルは画像データと BEV マップ両方の特徴マップを同じ構造のネットワークで計算する。ネットワークの前半は YOLOv3[2] を使う darknet-53 をベースに構築する。ネットワーク全ては 3x3 の畳み込み層と 11 層の残差ブロック（Residual）層で立ち上げた。総計 15 層で、深度 256

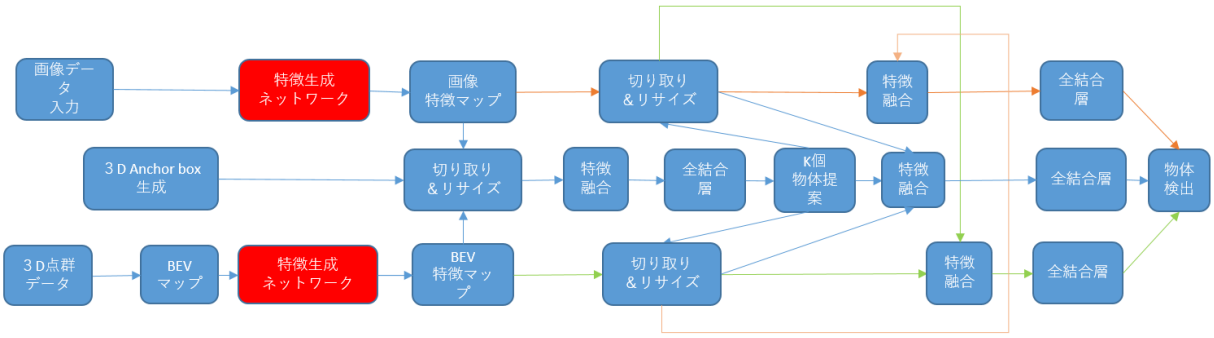

図 1　提案モデル

に立った。後半部分は、前半部分計算した特徴マップから、まず一回アップサンプリングし、前半部分同じサイズの特徴マップと融合する。次は、3x3 の畳み込み層で一回畳み込み計算を行い、1x1 畳み込み層から特徴マップを深度 32 に変換する。 1 つの特徴マップを得る。こうして、振り返し三つの特徴マップを生成する。これにより違う解像度の特徴マップからオブジェクトが検出易くなる。計算の負担も軽くなる。

2.2　マルチ特徴マップの切り取りとリサイズ変更操作と融合操作

[4] と同様に切り取りとサイズ変更操作で、特徴マップからすべてのアンカーの特徴クロップを抽出する。まず、3D のアンカーを BEV と画像特徴マップに投影することにより、二つの関心領域を得る。次に、対応する関心領域を使用し、画像特徴マップと BEV 特徴マップから各ビューの特徴クロップを抽出し、 3×3 にリサイズし、同じ長さの特徴ベクトルを得る。 最後、切り取りとリサイズサイズ変更操作した特徴マップは、要素ごとの平均操作（element-wise mean operation）で融合する。

2.3　損失関数

本研究提案モデルは二つのネットワークで構築される。まず一つ目のネットワーク区域提案ネットワークは提案アンカーを生成する。ネットワークの 3D ボックス回帰は Smooth L1 loss を使う、オブジェクトの損失はクロスエントロピー損失（cross-entropy loss）を使う。また、[1] と同じ様に、背景の回帰損失は無視する。 損失関数は [1] と同じ multi-task loss を使う。

$$L(pi, ti) = \frac{1}{N_{cls}} \sum L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum p_i^* L_{reg}(t_i, t_i^*) \qquad \text{式 (1)}$$

ここで、i はミニバッチ内のアンカーのインデックスであり、p_i はアンカー i がオブジェクトの確率である。グラウンドトゥルースラベル p_i^* は、アンカーが正であれば 1 であり、アンカーが負であれば 0 になる。 t_i は、アンカーボックス六つのパラメータ化された座標ベクトルであり、t_i^* は、正のアンカーに関連付けられたラウンドトゥルースボックスである。 分類損失 L_cls は、2 つのクラス（オブジェクトとオブジェクトではない）に対する log 損失である。$p_i^* L_{reg}$ は、回帰損失が正のアンカー $(p_i^* = 1)$ に対してのみアクティブになり、それ以外 $(p_i^* = 0)$ では無効になる。cls 層と reg 層の出力はそれぞれ pi と ti で構成されている。 λ はバランシングハイ

図 2　検出結果。青い枠は歩行者の検出結果、黄色はサイクリストの検出結果、赤い枠はグラウンドトゥルース。

パーパラメーターであり、重みを付ける。また、pi と ti は N_{cls} と N_{reg} によって正規化される。

二つ目のネットワーク検出ネットワークは、提案アンカーから違う尺度特徴マップからオブジェクトを検出する。複数の提案が BEV に同じスペースに回帰できるため、0.01 の IoU 閾値で 2D NMS を実行し、重複した検出を削除する。 損失関数は式 (1) と同じである。また、最終的な損失は区域提案ネットワークと検出ネットワークの損失総和である。

3　実験と結果

本稿は、特徴生成ネットワークの選択、3D アンカーのストライド (間隔) の選択、BEV 密度マップの選択と融合手法の選択について実験いた。

実験により特徴生成ネットワークは darknet-53 ベースの特徴生成ネットワークが一番良い結果が出た。検出精度と検出時間をバランスもつ上で、3D アンカーのストライドは 0.3m で選択した。BEV 密度マップは $\min\left(1.0, \frac{\log(N+1)}{\log(8)}\right)$ で生成した BEV 密度マップが一番良い結果が出た。最後に、モデルの大きさと計算コストを考え上で、融合手法を early 融合手法で選択した。前述より、特徴生成ネットワークを使い、early 融合手法、3 D アンカーのストライドを 0.3m で最後の結果は表 9 に示す。また、KITTI データセット実際の検出結果は図 2 に示す。青い枠は歩行者の検出結果で、黄色はサイクリストの検出結果、赤い枠はグラウンドトゥルースである。

本稿の実験は [4] が提供したグランドプレーンで 3D バウンディングボックスを生成する。しかし、彼らは KITTI データセットテストセットのグランドプレーンが提供していない。また、グランドプレーン生成手法も公開していないため、本稿では [4] と相似のグランドプレーン生成ため、3D 点群処理のための大規模ソフトウェアライブラリ Point Cloud Library (PCL) で生成した。グランドプレーンは最後 3 D バウンディングボックスの生成に影響されるため、テストセットの結果と期待の結果差があるになった。

4　結論と考察

本研究は AVOD をベースして、自動運転に向け新たなオブジェクト検出モデルを提案した。提案モデルは特徴生成ネットワークから画像と BEV 各三つ違う解像度の特徴マップを生成し、画像と BEV 特徴を融合し、三

	$3D_{AP}(\%)$			$BEV_{AP}(\%)$			
手法	easy	moderate	hard	easy	moderate	hard	runtime(s)
AVOD-FPN[4]	50.8	42.81	40.88	58.75	51.05	47.54	0.10
OHS-Direct [5]	51.29	44.81	41.13	55.90	49.48	45.79	0.03
TANet[6]	53.72	44.34	40.49	60.85	51.38	47.54	0.035
提案手法	35.99	27.77	25.19	44.50	35.15	25.19	0.21
AVOD-FPN(val)	48.90	46.29	41.61	54.55	50.24	46.80	0.11
提案手法 (val)	57.64	53.00	47.67	63.90	58.43	53.53	0.21

表 1　kitti test データセットに、IoU を 0 .5 で歩行者クラスの結果。val は検証（validation）データセットの結果。

つ融合した違う尺度特徴マップからオブジェクトを検出する。

KITTI データセットの実験により、提案モデルは他手法より歩行者検出の表現が優秀であることが分かった。しかし、AVOD のアーキテクチャから逃げていないので、また提案手法が小さいオブジェクト検出の方にファインチューニングし、大きいオブジェクト（例えば車）の検出結果はベースにした手法 AVOD とあんまり変わらなかった。

また、今回歩行者の検出方が良い結果でたか、サイクリストの検出はまだ問題がある。特に moderate 難易度と hard 難易度のオブジェクト検出は最新モデルよりまだ差がある。また、三つの特徴マップを使うため、三つの検出ネットワークを生成し、モデルのサイズがベース手法より 3 倍ぐらい大きくなっている。モデルの大きさを控え、moderate 難易度と hard 難易度のサイクリストの検出が今後課題になる。

最後、グランドプレーンの原因でテスト最終結果が期待と外したため、[4] らのグランドプレーン生成手法にアプローチするが、また新たな 3 D バウンディングボックスエンコーディング手法を探索するも今後課題である。

参考文献

[1] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems. 2015.

[2] Redmon, Joseph, and Ali Farhadi. "Yolov3: An incremental improvement." arXiv preprint arXiv:1804.02767 (2018).

[3] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia. Multi-view 3d object detection network for autonomous driving. In IEEE CVPR, 2017.

[4] Ku, Jason, et al. "Joint 3d proposal generation and object detection from view aggregation." 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018.

[5] Chen, Qi, et al. "Object as Hotspots: An Anchor-Free 3D Object Detection Approach via Firing of Hotspots." arXiv preprint arXiv:1912.12791 (2019).

[6] Liu, Zhe, et al. "TANet: Robust 3D Object Detection from Point Clouds with Triple Attention." arXiv preprint arXiv:1912.05163 (2019).