

# DROW: Real-Time Deep Learning based Wheelchair Detection in 2D Range Data

2022年6月16日 14:57

## DROW : 2Dスキャンデータでのリアルタイムディープラーニングベースの車椅子検出

<https://arxiv.org/abs/1603.02636>

<https://github.com/VisualComputingInstitute/DROW>

<https://arxiv.org/abs/1804.02463>

### Abstract

We introduce the DROW detector, a deep learning based detector for 2D range data. Laser scanners are lighting invariant, provide accurate range data, and typically cover a large field of view, making them interesting sensors for robotics applications. So far, research on detection in laser range data has been dominated by hand-crafted features and boosted classifiers, potentially losing performance due to suboptimal design choices. We propose a Convolutional Neural Network (CNN) based detector for this task. We show how to effectively apply CNNs for detection in 2D range data, and propose a depth preprocessing step and voting scheme that significantly improve CNN performance. We demonstrate our approach on wheelchairs and walkers, obtaining state of the art detection results. Apart from the training data, none of our design choices limits the detector to these two classes, though. We provide a ROS node for our detector and release our dataset containing 464k laser scans, out of which 24k were annotated.

2次元レンジデータで動作する深層学習ベースのオブジェクト検出器である DROW 検出器を紹介する。レーザースキャナは照明に左右されず、正確な2Dレンジデータを提供し、一般的に大きな視野をカバーするため、ロボティクスアプリケーションにとって興

味深いセンサーである。これまでのところ、レーザースキャナの2次元距離データの検出に関する研究は、手作業で作成した特徴量とブーストされた分類器が主流であり、最適でないデザインの選択により性能が低下する可能性がある。我々は、このタスクのためにConvolutional Neural Network (CNN)に基づく検出器を提案する。我々は、2次元レンジデータにおける検出のためにCNNを効果的に適用する方法を示し、CNNの性能を大幅に向上させる深度前処理ステップと投票スキームを提案する。我々は、車椅子と歩行器について我々のアプローチを実証し、最先端の検出結果を得ることができた。しかし、学習データを除けば、我々の設計上の選択のいずれも、検出器をこれら2つのクラスに限定するものではない。我々は、我々の検出器のためのROSノードを提供し、464kレーザースキャンを含む我々のデータセットを公開する（うち24kは注釈付き）

# I. INTRODUCTION

2022年6月16日 15:04

## 2Dレーザースキャナー

- 多くの自律型ロボットは2Dレーザースキャナーを搭載しており、一般的に人[1]、[34]や物体[18]の検出を含むナビゲーション関連のタスクに使用されている。
- レーザースキャナは、一般的に視野が広く、照明や環境条件に対して不変であるため、広く使用されている。

## 2Dレーザースキャナーによる検出

- 初期の検出方法は、直線や円のフィッティングのような単純なヒューリスティックを使用していたが[35]、過去数年間は、学習した分類器と結合した手作りの特徴が、レーザースキャナーの検出を支配してきた。
- このパラダイムの中で、モバイルロボットのナビゲーション[8]や自律走行[28]をサポートするために、成功した人物[1]、[30]、[20]、移動支援[34]、道路障害物[18]の検出器が幅広く開発されてきました。
- しかし、2次元距離データから得られる情報は、1回のスキャンで確実に検出するには不十分であり、センサフュージョン[30]、多層センサ設定[20]、[29]、トラッキングによる情報の時間統合に依存するアプローチが一般的であるようである。
- 我々は、単一の2Dレンジスキャンに基づく検出が実際に良好な性能を発揮することを示す。

## 本論文について

- 本論文では、特に車椅子と歩行者の検知に焦点を当てる。
- これは、高齢者介護施設において、多くの人々が移動補助器具を利用するサービスロボットの応用が動機となっている。
- 移動補助具の存在は、レーザースキャナーやカメラデータにおいて、人の外観を大きく変化させるため、既存の人物検出器では信頼性が低くなってしまう。
- しかし、特にこれらの補助具を使用している人は、近づいてくるロボットを避けるのが難しくなるため、信頼性の高い検出がより重要になる。
- [Weinrichら\[34\]](#)は同様の課題に直面し、個々のレーザースキャンから人、車椅子、歩行者を検出するためのGandalf検出器を提案している。

- 彼らはレーザーセグメントに対して新しい特徴セットを導入し、AdaBoost分類器を用いてこれらを分類する。
- その結果、Gandalf検出器は非常に良い性能を示したので、我々はこれをベースラインとして採用した。

## 本論文の手法について

- 本論文では、DROW 検出器を紹介する。
- 本論文では、車椅子と歩行者の検出に焦点を当て、その応用について述べる。
- しかし、我々のアプローチをこれらのオブジェクトに限定する設計上の決定はなく、十分な注釈付き学習データがあれば、人物や他のオブジェクトの検出にも一般化できると確信している。
- 我々の検出器の学習と実行に必要なコードとデータ、および新しいタスクのためにデータに注釈を付けるコードは、出版時に提供される予定です。

## CNNについて

- コンピュータビジョンにおいて、深層学習は最近新しいベストプラクティスとなり、手作業で作られた特徴を学習した特徴に置き換え、多くのタスクで技術の状態を一新しています[13], [32], [5]。
- 特に、畳み込みニューラルネットワーク(CNN)は挑戦的なタスクで非常に成功している。
- 本論文では、CNNをレーザーデータ中の物体検出に適用することで、特徴量エンジニアリングの必要性を軽減し、劇的な改善を可能にする方法を紹介する。

## 2DレーザースキャナーにおけるCNN

- MultiBox [16]やYOLO [24]のようなCNNベースの画像レベル検出器は原理的には2Dレーザースキャンに適用できるが、我々は素朴にそうすることが効果的でないことを発見した。
- レーザー点の空間密度は距離によって大きく変化するため、CNNの固定された知覚野は大きく異なるスケールをカバーし、学習が困難となる。
- レーザーセンサーが提供する空間情報を利用するために、我々は、各レーザー点の周りの一定の実世界の範囲の窓を切り出して正規化する前処理段階を提案する。
- これらのウィンドウは全てCNNに送られ、CNNはオブジェクトの位置に票を投じることが出来る。
- これらの票は、[Non-Maximum Suppression](#)を用いて個々の検出値に変換される。

我々は、深度前処理と投票の両方が我々のアプローチの重要な構成要素であることを示す（図1にその概要を示す）。

- 本アプローチは背景減算を必要とせず、シングルコアとGPUを用いた最新のデスクトップマシンにおいて、約75fpsのフレームレートで動作する。
- また、我々のロボットでは、ノートPCのGPUでレーザーのフレームレート（約13fps）に容易に追いつくことができました。

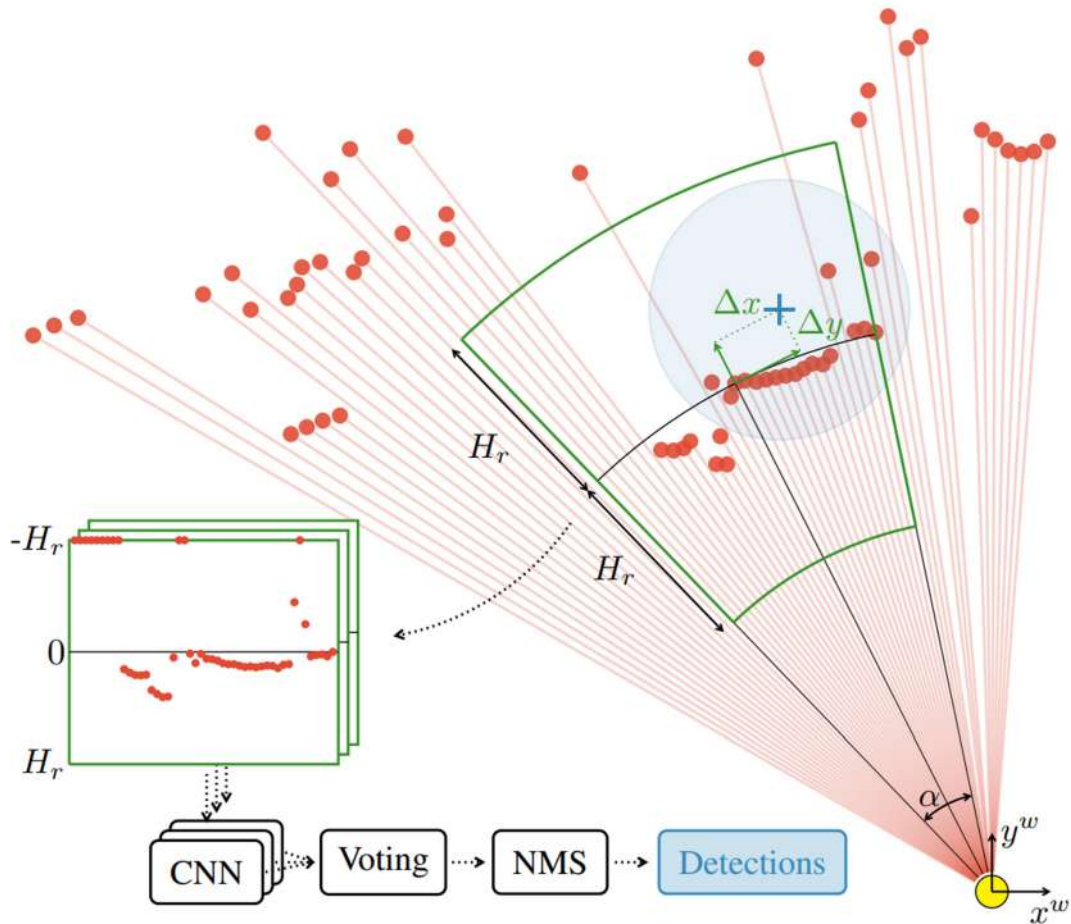


図1：本アプローチの概要。

レーザーの照射範囲に関係なく、現実の大きさが固定されたウィンドウを抽出する。

値はウィンドウの範囲にセンタリングされ、結びつけられる。

各ウィンドウはCNNによって分類され、オブジェクトが発見された場合、

その相対オフセット ( $\Delta x$ ,  $\Delta y$ ) に対して重み付けされた投票が行われる。

最後に、非最大抑制（NMS）を用いて投票が統合され、検出オブジェクトの重心が得られる。

## 要約

- DROWはCNNを用いた車椅子・歩行者検出器であり、2次元の距離データに対して、提供された奥行き情報を効果的に利用することで、最先端の結果を達成する。

- データセットは464kスキャンで、そのうち24kは車椅子と歩行者の中心位置が注記されている。
- 検出器のROSコンポーネントと、学習済みモデルを含む関連サービスモジュールを提供します。

## II. APPROACH

2022年6月16日 15:37

### 手法の流れ

- 前処理
  - 各レーザー点の周囲に再サンプリングされたウィンドウを切り出し、ローカル座標系での検出位置を計算する
- CNN
  - ウィンドウを分類し、相対的な検出位置を予測する。
- 統合
  - 投票と非最大抑制方式により、予測を検出に変換する。

### A. 2次元レンジデータの事前処理

#### 2次元スキャンデータでのCNNの課題

- CNNを適用して検出するためには、ネットワークの受容野が対象物の大部分を覆っている必要がある。
- レーザースキャンの問題点は、近くの物体には大量のレーザービームが当たるが、遠くの物体にはほんの一握りのビームしか当たらないことである。
- つまり、CNNの受光野はレーザーの大部分をカバーしなければならず、学習シーンの背景に非常にオーバーフィットしやすくなってしまう。

#### 2次元スキャンデータでのCNNの課題の解決策

- この問題を回避し、同時にレーザーデータが提供する実世界のスケール情報を利用するために、我々は深度型スライドウィンドウ方式(depth-guided slidingwindow fashion)でCNNを評価することを提案する。
- つまり、どの距離でも物体がほぼ同じ表現を持つようにデータを前処理することで、距離によって全く異なる表現を暗黙的に学習する必要をなくす。
- 各ビームの周りに、実世界の範囲 $l$ の窓を切り出す。
- したがって、角度 $\alpha = 2 \sin^{-1}(\frac{l}{2r})$ にまたがり、現在のビームが障害物に衝突する距離 $r$ に依存する可変量の測定値を含む。
- そして、このウィンドウ内の測定値を $n$ 個の固定サンプルで線形に再サンプリングする。
- このようなウィンドウに適用すると、ネットワークの受容野は距離 $r$ に関係なく常に同じ実世界の範囲をカバーする。
- さらに、ウィンドウを現在の点の周りに中心を置き、遠くの乱雑さを除去するために $\pm H_r$ の範囲外の値を留め、最後に値を $[-1, 1]$ に投影する。
- つまり、図1に示すように、深さ $r$ の点の周りのウィンドウの各点 $x$ に $\max(-H_r, \min(x - r, H_r)) / H_r$ が適用されることになる。
- これらの前処理操作のうち、どれがDROWの性能に最も寄与しているかについての詳細な分析は、セクションIII-Eに記載されている。



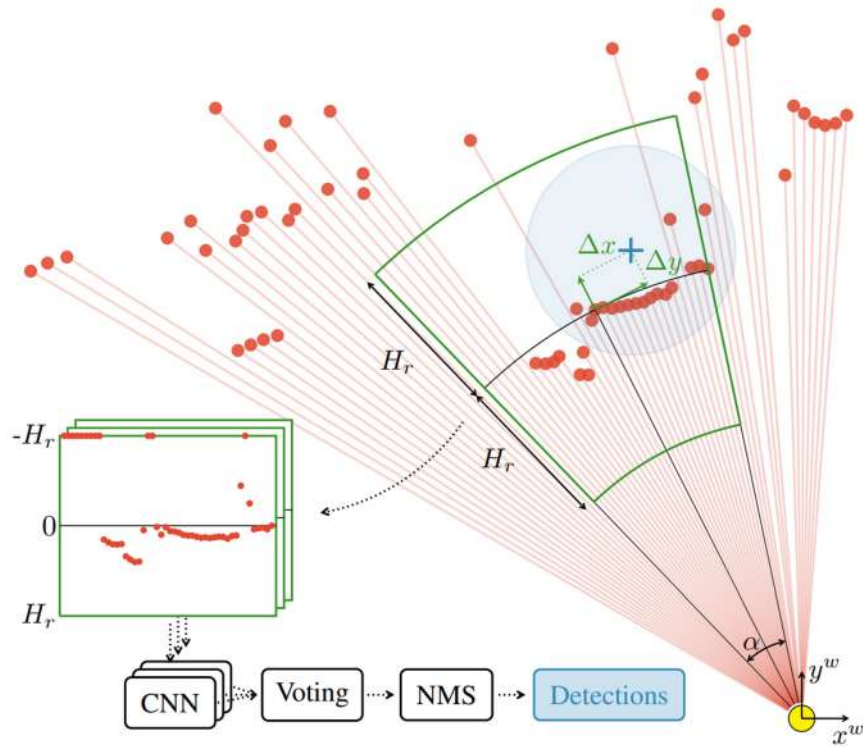


図1：本アプローチの概要。

レーザの照射範囲に関係なく、現実の大きさが固定されたウィンドウを抽出する。

値はウィンドウの範囲にセンタリングされ、結びつけられる。

各ウィンドウはCNNによって分類され、オブジェクトが発見された場合、

その相対オフセット（ $\Delta x$ 、 $\Delta y$ ）に対して重み付けされた投票が行われる。

最後に、非最大抑制（NMS）を用いて投票が統合され、検出オブジェクトの重心が得られる。

## B. 予測

- 各ウィンドウ内の各レーザポイントについて、CNNはSoftMax出力によりこのウィンドウが関心のあるオブジェクトクラスに属するかどうかを分類する
- もし物体なら、回帰出力によりそのオブジェクトの中心位置に投票する。
- 2次元レンジデータは本質的に回転不変であるため、絶対座標（ $x$ ,  $y$ ）で投票を行うことはしない。
- その代わりに、図1に示すように、現在のレーザ一点を中心とした座標系でオフセット（ $\Delta x$ ,  $\Delta y$ ）を学習する。

## C. 投票と非最大抑制（NMS）

- 各ウィンドウの予測は、検出センターに統合される必要がある。
- これは、CNNの予測値をレーザの視野にまたがる規則的なグリッドに投票させることで達成される。
- $p(0|w) = \sum_{c \in C} p(c|w)$  をウィンドウ $w$ が注目物体を見る総確率とし、 $c$ を検出すべきクラスとする。
- $p(0|w)$ があらかじめ定義された投票閾値 $\tau$ を超えた場合、ウィンドウは重み $p(0|w)$ を持つクラス非依存グリッドに投票するとともに、重み $p(c|w)$ を持つ各クラス特定グリッドにも投票する。
- すべてのウィンドウが潜在的に投票した後、各グリッドはガウシアンフィルタでぼかされ、クラス不可知論的グリッドでNMSが実行される。
- そのグリッドで見つかった各最大値に対して、当該セルで最も高い投票総数を持つクラスを使用して、セルの中心で検出が予測される。
- このような投票方式を採用した理由は、各クラスを別々に扱うのではなく、同じ位置で両方のクラスが検出されるのを避けるためである。
- 図2に各ステップの例を示す。(a)の生のレーザ点から投票されたものを(b)に、(c)-(e)は3つの投票グリッド、(f)は結果として



得られた2つの検出を示す。

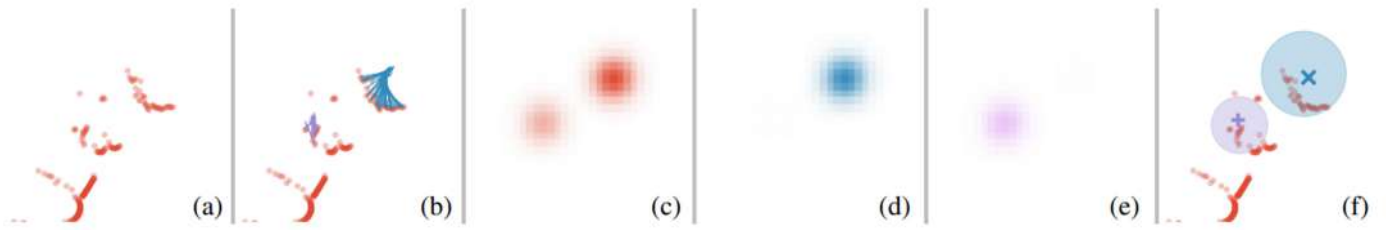


図2：非最大抑制（NMS）

(a)入カスキャンの一部

(b) 矢印で示された投票

(c)共同投票グリッド

(d)車椅子の投票グリッド

(e)歩行器投票グリッド

(f) 検出結果

# III. EXPERIMENTAL EVALUATION

2022年6月16日 17:55

我々のアプローチを評価するために、まず新しいデータセットを紹介する。そして、学習方法と評価方法の詳細を説明する。DROWの一般的な評価とベースラインとの比較の後、我々の検出器の個々の部分が全体の性能にどれだけ貢献しているかを示すために、いくつかのアブレーションスタディを実行します。

## A. データセット

- Weinrichら[34]はデータセットを提供しているが、彼らの記録は単一の車椅子や歩行器のあるシーンに限定されている。
- これは、かなり制約のある特徴量と分類器のいくつかのパラメータを学習するには十分かもしれないが、我々はゼロから特徴量を学習したいので、より一般的で多様なデータが必要である。
- そこで、ある高齢者介護施設で10時間強のデータを収録しました

### 1) 記録のセットアップ

- SCITOS G5ロボットにSICK S300レーザースキャナーを搭載し、地上から約37cmの高さに設置してデータを記録した。
- レーザーはほぼ13Hzで記録され、 $0.5^\circ$ の分解能で $225^\circ$ の視野を持ち、合計450回の計測を行うよう設定された。
- また、アノテーションのために、ロボットの頭部に取り付けたASUS XtionからRGB-Dデータを記録した（図3参照）。
- 残念ながら、プライバシー保護の観点から、ビデオストリームを公開することはできません
- ソフトウェア基盤はROSをベースとし、すべてのデータはrosbagsに保存されました。
- 車椅子と歩行器を十分に見ることができるように、ロボットが録画している間、一人の人間がどちらかを使って常時歩き回っていました。
- 記録は、(1) 植木鉢、椅子、家具、ローリングベッドなど、あらゆる雑多なものを含む施設内の日常生活を記録した自然風景と、(2) 人工的な記録の両方から構成されています。
- (2)データセットにバリエーションを持たせるために、ある一定のパターンで車を走らせた人工的な記録である。
- また、入居者の車椅子とは別に、電動車椅子1台を含むできるだけ多くの車椅子・歩行器モデルを収録した。



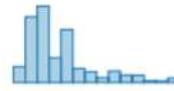
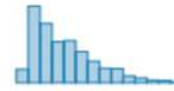






図3：開発したロボット「カール」

## 2) データセットの統計

- 得られたデータセットを、トレーニングセット、検証セット、テストセットに分割しました。
- これらのセットを作成するために、介護施設を4つの重複しないエリアに分割し、そのうちの3つをそれぞれ訓練セット、テストセット、検証セットに割り当てた。
- 4つ目のエントランスホールは、時間的に不連続なシーケンスに分割され、訓練セットと検証セットに分配された。
- この分割に基づき、本アプローチが未知の領域に対してどの程度汎化できるかを示すことができる。
- 表Iは我々のデータセットの概要と、アノテーションを行ったサブセットの統計情報を示している。
- 車椅子と歩行者のカウントは、インスタンスではなく、個々の検出を意味し、棒グラフは距離の分布を示す。
- 各棒は距離の1mスライスを表し（15mまでは15本）、観測された移動補助器具の大部分がロボットから1m～6m以内で遭遇していることが明確に示されている。

表1: データセット概要

	Train	Validation	Test	Total
Sequences	78	30	5	113
Scans	341 138	74 744	48 131	464 013
Annotated Scans	17 665	3 919	2 428	24 012
Wheelchairs	14 455	5 595	1 970	22 020
Walkers	2 047	219	581	2 847
Wheelchairs by distance				
Walkers by distance				

## 3) アノテーション

- 車いすと歩行者の中心をアノテーションする。
- データセットには464k枚のレーザースキャンの生データが含まれているため、作業量を管理しつつ、シーケンスの全範囲をカバーし、将来的な時間的アプローチの開発を可能にするアノテーション方式を考案した。
- 全てのレーザースキャンをアノテーションするのではなく、以下のように各シーケンスに小さなバッチをアノテーションする。
- 1バッチは100フレームで構成され、そのうち5フレームごとにアノテーションを行うため、1バッチあたり20フレームがアノテーションされる。
- このツールは、スキャンシーケンスと頭部カメラからの対応するRGB画像をロードし、アノテーションが必要なバッチを自動的に検出する。
- このツールは、頭部カメラからのRGB画像に対応したスキャンシーケンスを読み込み、アノテーションが必要なバッチを自動的に見つける。
- アノテーションを支援するために、現在のバッチの最初の画像、現在の画像、最後の画像を表示させる。
- このアノテーションツールの表示例として、図4をご覧ください。
- マウスポインタの周囲には、車椅子の平均的な大きさを示す1.2mの円が表示され、車椅子や歩行器の中心位置をクリックしやすくなっています。
- このような情報を活用することで、ほとんどのモビリティエイドにアノテーションを施すことができましたが、限られた

カメラ視野の中で、いくつかのモビリティエイドを見落としたと思われます。

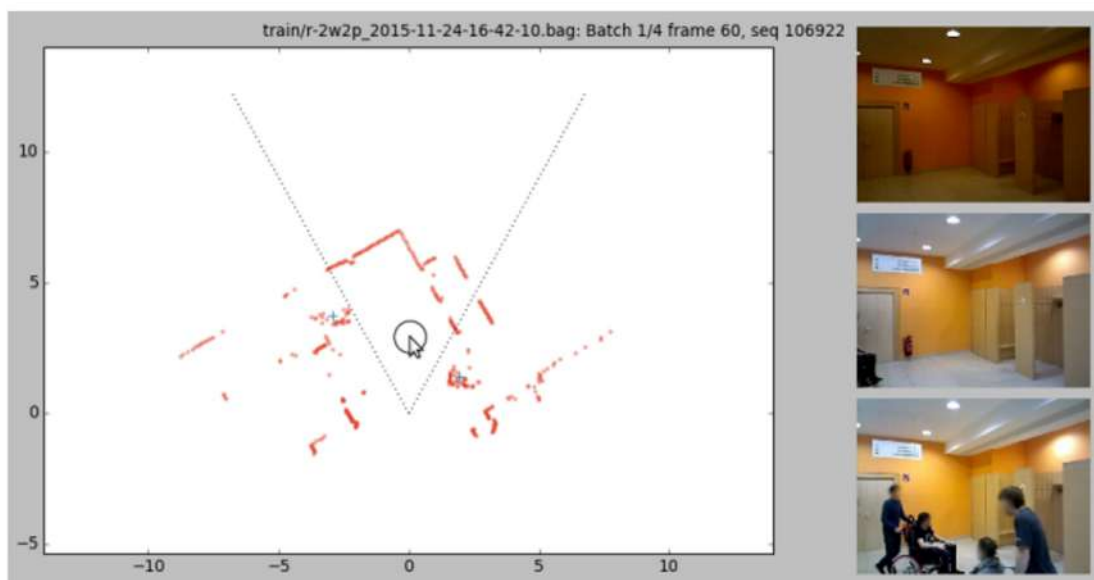


図4: アノテーションツールの表示例。

点線の円錐はXtionの視野を示し、

右側にはバッチの最初の画像、現在の画像、最後の画像が表示されている。

青い十字はアノテーションされた車椅子を示している。

## B. トレーニングの手順

- 我々は、前処理された各窓について、それが近くの車椅子や歩行者を示しているかどうかを予測し、もしそうであれば、その中心のオフセットを予測するCNNを学習する。
- これは、背景、車椅子、歩行者を区別するthree-way SoftMaxと、2次元線形回帰出力の2つの出力層を持つネットワークを用いて、1回のパスで行うものである。
- 我々は、SoftMax出力における負の対数尤度基準と回帰出力における二乗平均平方根誤差の和を最小化することによって、ネットワークを最適化する。
- 回帰対象は図1に示すように各ウィンドウのローカル座標系で $\Delta x$ ,  $\Delta y$ として計算される。
- クラスラベルはウィンドウの中心点に最も近い検出の種類によって決定され、最大ユークリッド距離は車椅子で0.6 m、歩行者で0.4 mである。
- 近傍にラベルがない場合、回帰出力に対して誤差はバックプロパゲートされず、ネットワークは任意のオフセットを自由に予測することができる。
- このネットワークのアーキテクチャは、一般的なVGGnet [27] に触発されたもので、以下の通りである。

1. Conv 5@64
2. Conv 5@64
3. Max 2
4. Conv 5@128
5. Conv 3@128
6. Max 2
7. Conv 5@256
8. Conv 3@5

- ここで Conv  $n@c$  はサイズ $n$ の $c$ 個のフィルターによる畳み込み
- Max  $p$ は $p$ 個の値のウィンドウに対する最大の演算を表す。
- バッチ正規化[12]
- 0.25のドロップアウト[31]
- およびReLU非線形[9]がすべての畳み込み層の間に適用される。

レイヤー	処理	入力チャンネル数	入力幅	入力高さ	パディング	ストライド	フィルタサイズ	プーリング幅	プーリング高さ	出力チャンネル数	出力幅	出力高さ	ニューロン数
1	畳み込み層	1	48	1	0	1	5			64	44	1	2816
2	畳み込み層	64	44	1	0	1	5			64	40	1	2560
3	プーリング層	64	40	1				2	2	64	20	1	1280
4	畳み込み層	64	20	1	0	1	5			128	16	1	2048
5	畳み込み層	128	16	1	0	1	3			128	14	1	1792
6	プーリング層	128	14	1				2	2	128	7	1	896
7	畳み込み層	128	7	1	0	1	5			256	3	1	768
8	畳み込み層	256	3	1	0	1	3			5	1	1	5

- 注目点前後48個にリサンプルされた入力窓に対して、ネットワークは長さ5のベクトルを出力し、そのうちの3つはSoftMaxに送られ、残りの2つは回帰出力 $\Delta x$ と $\Delta y$ となる。
- ネットワークの学習には、AdaDeltaオプティマイザーを用い、 $\rho = 0.95$  と  $\epsilon = 10^{-7}$  で、近似的に収束するまで学習する。
- 学習中、回帰目標に小さな乗法ランダムノイズを加え、各窓とその目標を確率0.5で反転させる。
- このCNNはTheano[2]で実装した。
- 投票方法としては、投票に乗法的なクラス重みを加えて正規化し、

$p(c|w)$ の代わりに

$$\frac{w_c p(c|w)}{\sum_{c \in C} w_i p(i|w)}$$

を用いてクラス $c$ に投票するようにする。

- 次に、検証集合において $\max_T f_{\text{wheelchair}}(T) + f_{\text{walker}}(T)$  ( $f_c(T)$ は検出閾値 $T$ を用いたクラス $c$ のF1スコア)を最大化するためにhyperopt [4] を用いて全ての投票ハイパーパラメータ (クラス重み $w$ 、グリッド解像度 $b$ 、ブラーサイズ $\sigma$ ) を最適化する
- 興味深いことに、 $w_{bg} = 0.38$ ,  $w_{\text{wheelchair}} = 0.60$ ,  $w_{\text{walker}} = 0.49$ ,  $b = 0.1 m$ ,  $\sigma = 2.93$  という最適値は、我々の初期推測から大きく外れていない。

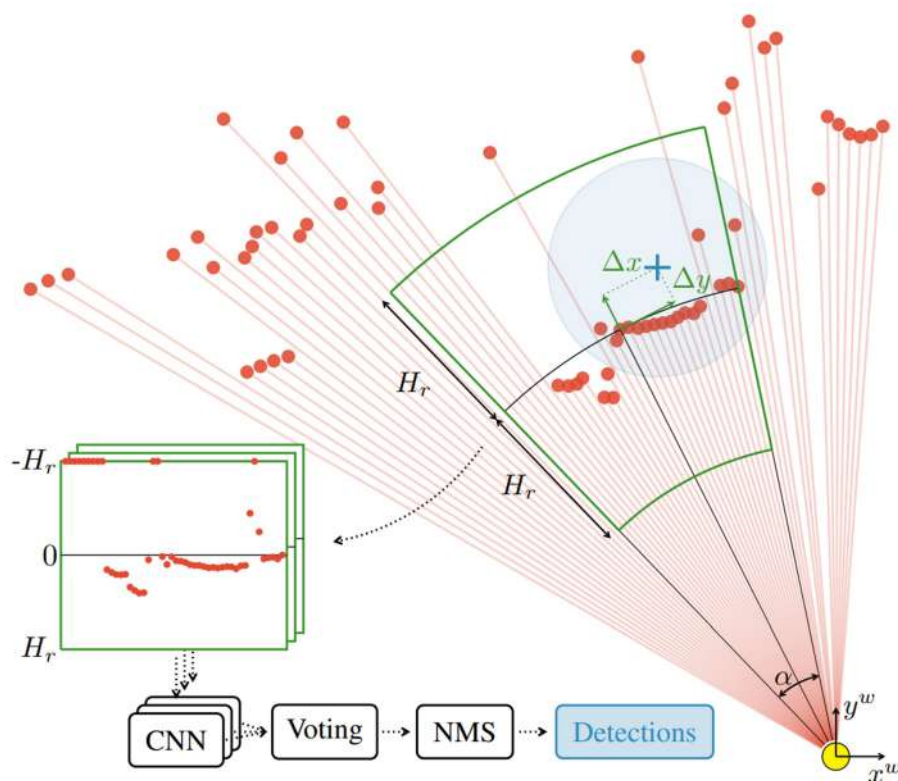


図1：本アプローチの概要。

レーザーの照射範囲に関係なく、現実の大きさが固定されたウィンドウを抽出する。

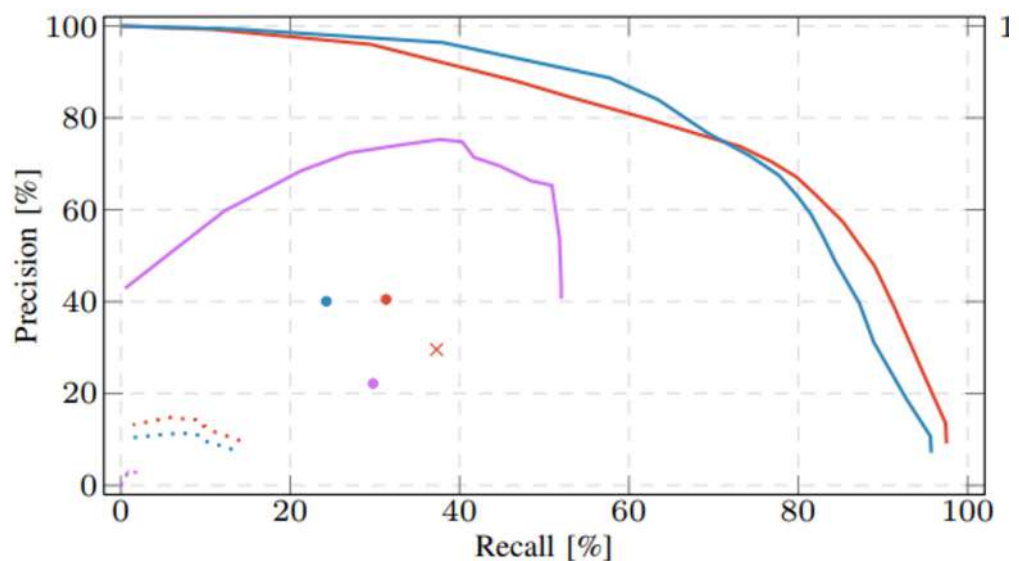
値はウィンドウの範囲にセンタリングされ、結びつけられる。



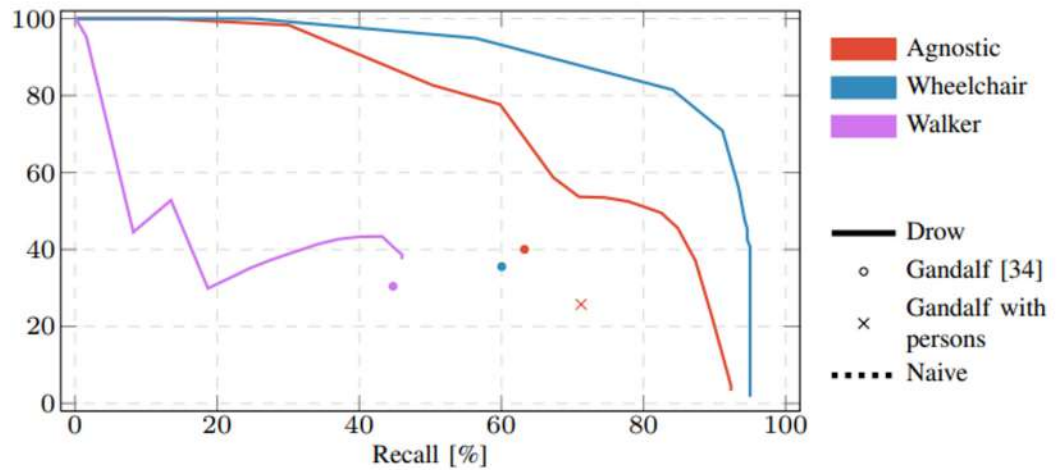
各ウィンドウはCNNによって分類され、オブジェクトが発見された場合、その相対オフセット ( $\Delta x$ ,  $\Delta y$ ) に対して重み付けされた投票が行われる。最後に、非最大抑制 (NMS) を用いて投票が統合され、検出オブジェクトの重心が得られる。

### C. アプローチ評価

- 我々は、前のセクションで説明したように、学習セットでモデルを学習し、検証セットでハイパーパラメータを計算しました。
- DROWの実世界での性能を評価するために、我々のテストセットと、同様のロボットで記録された一般に公開されている [34] のRehaテストセットの精度曲線と再現曲線を見ます。
- 我々のテストセットは介護施設内の見たことのない場所で記録されているため、位置の事前予測や背景モデルを学習する方法は失敗することを想起してください。
- 図5に示す結果は、DROWが非常によく汎化し、我々のテストセットとリハのテストセットの両方で、Gandalf [34]よりも有意に高い精度と再現率を持つことを示しています。
- 我々のアプリケーションシナリオでは、モビリティエイドを正しく分類することよりも、実際に検出することが重要であるため、クラスに依存しない性能も評価する。
- このため、評価時にはすべての検出と注釈のクラスを無視する。
- 評価半径は0.5mとし、検出が真実の中心から0.5m以内にあり、正しいクラスを持つ場合、その検出は真実とマッチングすることを意味する。
- 注釈は最大1つの検出とマッチし、そのクラスの残りの検出はすべて偽陽性、そのクラスのマッチしない注釈はすべて偽陰性である。



(a) テストセットでの性能



(b) Rehaテストセット[34]に対する性能。

図5：2つのテストセット (a) 、 (b) におけるDROW、Gandalf検出器[34]、  
ナイーブな深層学習ベースラインの性能比較。

セクション III-D で説明したように、

ナイーブなベースラインは Reha テストセットでは適用できない。

## 混合行列

クラス分類の予測結果が下記の4つの種類に分けられます。分かりやすくするためにウイルス感染検査手法の精度検証課題を例に挙げて説明します。

- **真陽性 (True Positive: TP)** : ポジティブサンプルに対して正しくポジティブと予測されたサンプル数  
例) ウイルス感染者の中に、正しく陽性と判定された人数
- **偽陰性 (False Negative: FN)** : ポジティブサンプルに対して間違ってネガティブと予測されたサンプル数  
例) ウイルス感染者の中に、陰性と判定された人数
- **真陰性 (True Negative: TN)** : ネガティブサンプルに対して正しくネガティブと予測されたサンプル数  
例) ウイルス感染していない人の中に、正しく陰性と判定された人数
- **偽陽性 (False Positive: FP)** : ネガティブサンプルに対して間違ってポジティブと予測されたサンプル数  
例) ウイルス感染していない人の中に、陽性と判定された人数

上記の情報を下記のように混合行列でまとめられます。

	予測結果が 陽性の場合	予測結果が 陰性の場合
正解が 陽性の場合	真陽性 (TP)	偽陰性 (FN)
正解が 陰性の場合	偽陽性 (FP)	真陰性 (TN)



## 適合率 (Precision)

適合率とは、ポジティブクラスと予測したサンプルの中に、どのくらい正しく予測できたかの割合です。予測結果が陽性になったものを注目しています。ネガティブサンプルの誤認識 (FP) が多いほど、適合率が低くなります。以下の式で計算されます。

$$Precision = \frac{TP}{TP + FP}$$

### いつ使うべきか：

誤認識・誤検知をなるべく抑えたいとき。例えば、迷惑メール判定課題では正しく迷惑メールを認識できること (TP) は大事ですが、重要なメールの見逃しが発生すると困りますね。そのため、迷惑メールではないものの誤検知 (FP) をなるべく抑える必要があります。

## 再現率 (Recall)

再現率とは、ポジティブクラスの中にどのくらい正しく予測できたかの割合です。正解が陽性になったものを注目しています。ポジティブサンプルの見逃し (FN) が多いほど、再現率が低くなります。以下の式で計算されます。

$$Recall = \frac{TP}{TP + FN}$$

### いつ使うべきか：

見逃しをなるべく抑えたいとき。例えば、障害や異常検知課題などです。機械の故障が生じると、事故や損害を発生するリスクがあるため、このような課題ではなるべく見逃しを抑える必要があります。しかし、正常時に頻繁に間違えて異常として検知されると、そのシステムが使い物にならないですね。そのため、再現率を優先された課題でも、再現率が高いほど良いとは限らず、適合率も許せるレベルまで達成できたかを確認する必要があります。

<https://tech-blog.optim.co.jp/entry/2021/05/31/100000#%E5%86%8D%E7%8F%BE%E7%8E%87Recall>

- DROWの検出がどの程度局所化されているかを見るために、評価半径を変化させたときの精度-再現曲線を図6に示します。
- 半径0.3m以上では、すべての曲線が乱雑になり、DROWの検出がうまく局所化されていることがわかります。
- アノテーションの際には、アノテーションの精度に限界があることが明らかになりました。つまり、半径0.1mという小さな半径での評価は、ラベリングノイズの影響を強く受けることがわかります。
- また、レーザースキャナーからの距離に対するDROWの挙動を分析しました。
- まず、レーザーから0.1m以上の距離では、すべての検出を無視することから始めます。

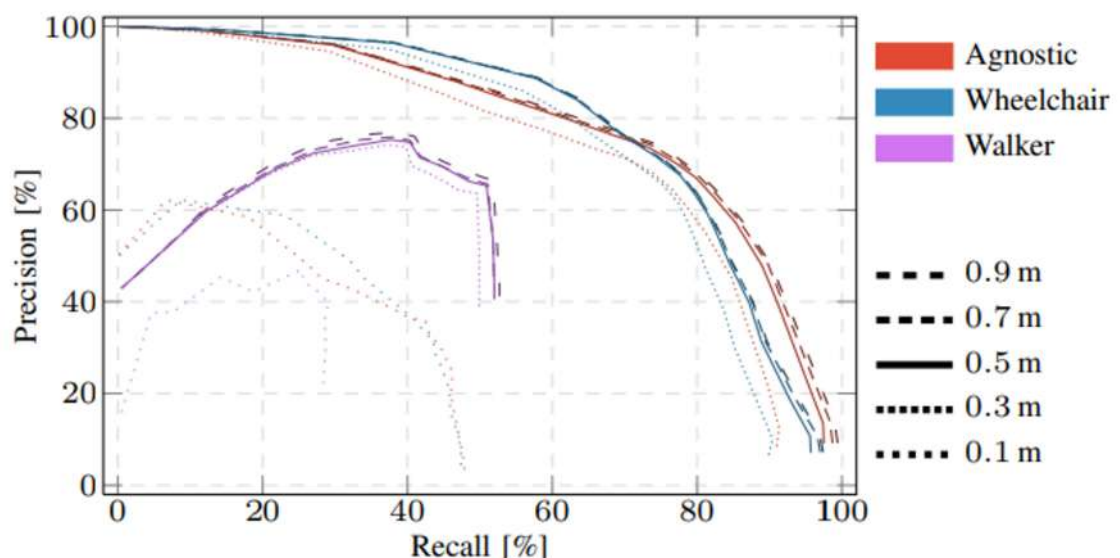


図6：異なる評価半径に対するDROWの性能。

0.3以上の半径で高いオーバーラップを示していることから、

我々の予測がうまく局在していることがわかる。

- そして、徐々に半径を広げ、より多くの検出を考慮に入れ、一定の閾値における精度と再現率がどのように変化するかをプロットしたのが図7です。
- DROWは全体的に非常に優れた性能を発揮しますが、特にナビゲーションとプランニングに重要な中間距離でその性能が発揮されます。
- 当然のことながら、10mの距離を超えると曲線はあまり変化しません。
- なぜなら、その距離まで観測された移動補助装置はごくわずかだったからです。

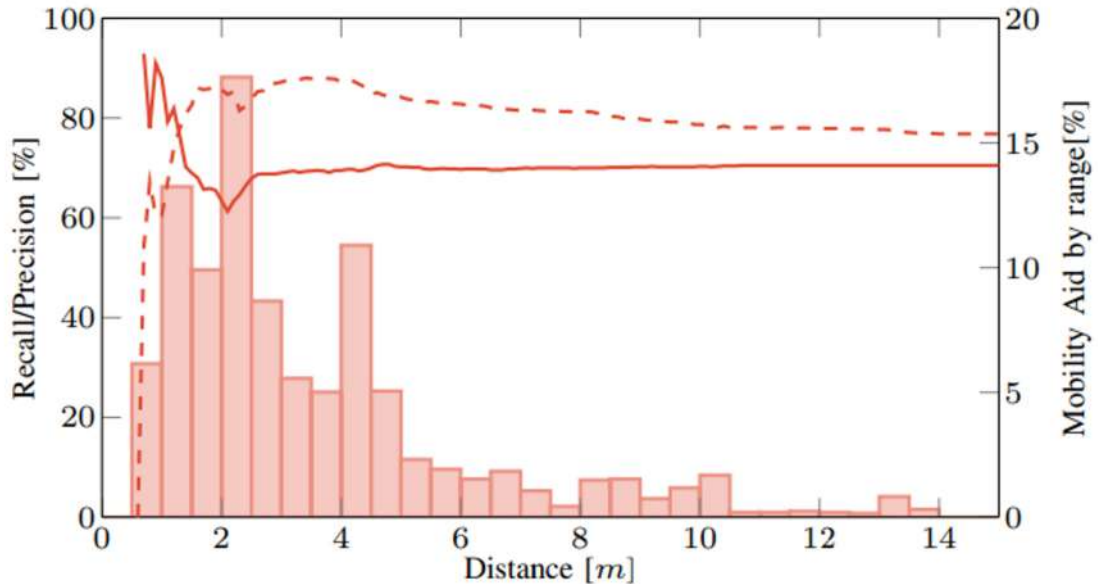


図7：T=0.5におけるクラス不可知論的な場合の、

ある距離（m）までの精度（— %）と再現率（--- %）。

ヒストグラム（■、右のy軸）は、ある範囲での全アノテーションの割合を示している。

## D. ベースライン

- 図5で提案手法を2つのベースラインと比較します：
  - 一般に公開されているGandalf検出器[34]とナイーブなディープラーニングのベースラインです。
  - 精度-再現曲線はいずれもGandalfの実際の検出性能を定量化しておらず、システムの一部に独立して着目しているに過ぎません。
  - 比較可能な検出精度-再現性の値を得るために、上記の評価プロトコルを使用してGandalf5を評価します。
  - Gandalfは閾値を調整する必要がないため、プロット上では1点となっています。
  - また、クラス分けをしない場合、Gandalfの人物検出を残す場合（・）と捨てる場合（×）の結果をそれぞれプロットし、精度とリコールのトレードオフを示す。
- [34]のコードを用いて得られた検出器の性能は、元の論文で報告されたものよりも大幅に低いことに注意してください。
- この性能は[34]で簡単に言及された共分散ベースのマージステップによって改善される可能性があります、このステップは論文で詳しく説明されておらず、また提供された検出器のコードにも含まれていません。
- このため、[34]で提示された結果を再現することはできませんでした。
- しかし、提供されたコードを使用して観察された誤検出と欠落検出から推定すると、この方法の性能に関する楽観的な境界でさえ、我々の検出器の曲線より下に置かれるでしょう。
- ナイーブな深層学習のベースラインとして、YOLO[24]と同様に、フルスキャンに基づき、正規化（x, y）座標で最大2つ

の検出を直接予測する別のCNNを訓練する（フレームの95.0 %に十分）。

- このベースライン実験の結果は図5（a）の点線で示されているが、Rehaテストセットではスキンのレーザー光量が異なるため、欠落している。
- 注意すべきは、我々はナイーブCNNのベースラインをできるだけうまく動作させ、ベストプラクティスに従うようにかなりの時間を費やしたことである：
  - 慎重に選ばれた受容体、バッチ正規化、慎重な初期化、AdaDeltaオプティマイザ、などなど。
  - しかし、簡単にわかるように、このような素朴な方法で2Dレンジデータに深層学習を適用しただけでは、ひどい結果になります。

## E. アブレーション研究

- ここでは、我々のアプローチを定義する各パーツを系統的に削除または置換することで、我々の設計上の決定が検出にどのような影響を与えるかを分析する。
- 図8は、これらの実験をまとめたものである。
- 全体として、完全なDROW検出器（一）は、他のすべての構成より優れている、つまり、それぞれの単一の決定が高い性能に寄与している。
- 以下の各実験では、完全なネットワークをゼロから再トレーニングする。

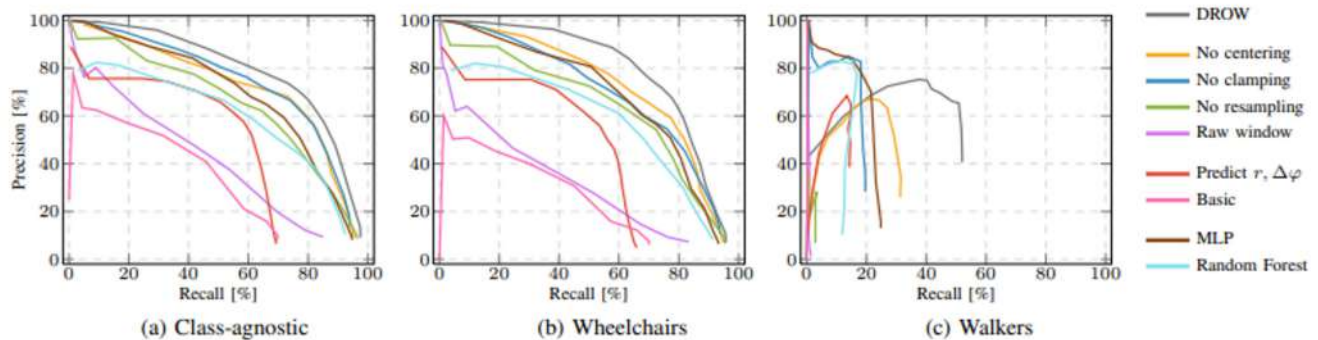


図8：アブレーションの研究。パフォーマンスがどのように変化するかを示すために、我々のアプローチに若干の修正を加えて再トレーニングを行った。

### 1) 前処理

- 深度の前処理は、センタリング、クランピング、リサンプリングの3つのオペレーションに分解されます。
- 入力ウィンドウをセンタリングしないことは、無視できないものの、最も小さな効果をもたらします。
- クランプもセンタリングと同様、全体的なパフォーマンスを低下させる。
- ネットワークの受容野が一定の実世界サイズを保つことを保証するリサンプリングステップがないと、特にウォーカーに対して性能が著しく低下する。
- CNNが異なる距離で全く異なる表現を学習しなければならないことを考慮すると、全体的な性能はまだ驚くほど良好である。
- 歩行者がほとんど検出されず、なおかつ不可知論的な性能がまともであることは、高い混同を示唆している。
- 最後に、上記のステップを全て削除した場合、全ての指標において最悪の結果となり、我々の前処理が実に重要であることが示された。
- これらの実験から、3つの前処理がそれぞれ重要であることが明らかになった。

### 2) 投票

- 図5aの素朴な深層学習ベースラインとの比較で、適切な検出結果を得るためには投票が不可欠であることを既に示しまし

た。

- ローカルオフセット ( $\Delta x$ ,  $\Delta y$ ) を回帰する効果を見るために、我々は ( $\Delta \phi$ ,  $r$ ) : 角度オフセットとレーザースキャナへの絶対距離を回帰するネットワークのバージョンを訓練します。
- この実験では、センタリング (c.f.) も除去する必要がある、そうしないと  $r$  が予測できないからです。
- センタリングがなければ、ネットワークは原理的に中心点の距離を「通過」するように学習し、最後の層でバイアスとして追加することができる。
- しかし、この結果は、この学習方法が解決できない、より困難な問題であることを示唆している。
- この困難さが我々の前処理との悪い相互作用によって引き起こされていないことを確認するために、我々はまた、未処理の生のデータに対してこれらの予測を行うネットワークを訓練する。
- このモデルの性能は最悪で、平凡な投票ベースのネットワークだけでも、入力と出力の両方に注意を払わないと、うまくいかないことがわかる。

### 3) モデル

- すべての窓は個々にモデルを通して送られ、すべての窓はリサンプリングステップのために同じサイズであるため、モデルがCNNである固有の理由はない。
- CNNにエンコードされた空間事前情報が有用であることを示すために、我々は3つの隠れ層パーセプトロンを、それぞれ2048の隠れユニット、ReLU非線形性、ドロップアウト、バッチ正規化で訓練する。
- 他の実験と同様に、他の設計上の決定はすべて変更しないままとした。
- その結果、MLPは有力な代替手段であるが、CNNは明らかにそれを上回っていることがわかった。
- 追加のベースラインとして、3つのクラス確率とローカルオフセット ( $\Delta x$ ,  $\Delta y$ ) を回帰する回帰林をトレーニングしました。
- これはCNNとMLP（フリップ拡張を含む）と同じ訓練データで訓練されました。
- 回帰の森はscikit-learn[23]の実装を使用し、すべてのデフォルト設定と50本の木を使用しました。
- トレーニングは7.5時間かかり（5スレッドを使用）、我々のCNNモデルの1.1MBと比較して、6.9GBのモデルが得られました。
- それにもかかわらず、フォレストモデルはすべてのケースでMLPよりもさらにわずかに悪いパフォーマンスを示しています。
- 窓の大きさ 車椅子の幅が1.2mであることから、1.66m幅のウィンドウで評価した。
- この実験の結果は、図9に示すとおりである。この結果は、入力窓の大きさを複数用意することで、さらなる改善が期待できることを示唆している。

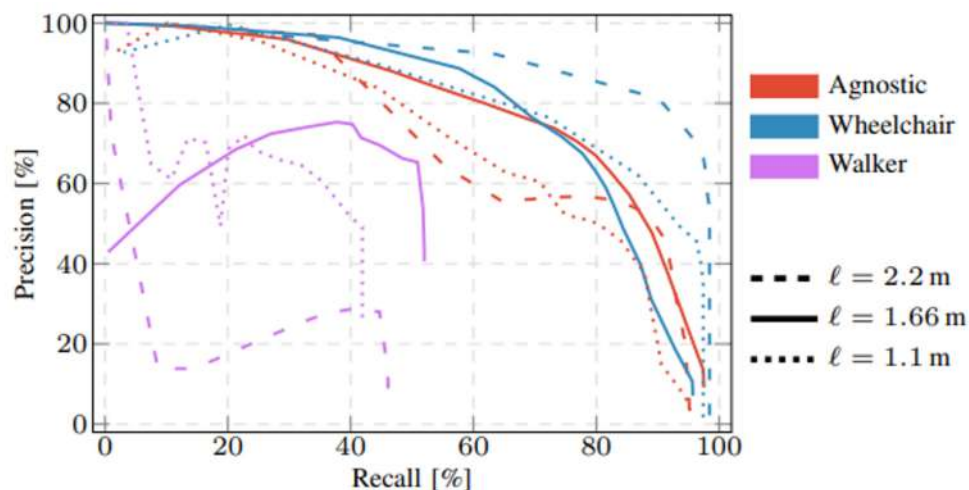


図9：ウィンドウサイズが性能に与える影響

#### 4) 窓の大きさ

- これまで、車椅子の長さが1.2mであることから、 $l=1.66\text{m}$ の窓で評価を行ってきたが、この選択が妥当であることを確認するため、より大きな窓（2.20m）と小さな窓（1.10m）でネットワークを学習させた。
- この実験の結果は、図9に示すとおりである。
- この結果は、入力窓の大きさを複数用意することで、さらなる改善が期待できることを示唆している。

## IV. RELATED WORK

2022年6月16日 20:19

- 我々のアプローチに最も関連するのは、Weinrichら[34]のアプローチであり、彼らはレーザーベースの検出のための距離不変の特徴を提案している。
  - 彼らは、まず距離の大きなジャンプを検出することによって、2次元レンジデータから車椅子、歩行者、そして人を検出する。
  - このようなジャンプが発見されると、彼らはその後のスキャン点をカバーする一定の実世界範囲を持つウィンドウを作成する。
  - この窓は等しい大きさのセグメントに分割され、それぞれのセグメントに対して、窓の深さに対するクランプされた距離が計算される。
  - 各セグメントは、最小、最大、平均の深さによって特徴付けられ、すべてのセグメントを連結して特徴ベクトルが形成される。
  - そして、AdaBoost 分類器によって分類される。
  - この特徴量設計は、我々の深度前処理ステップと高い類似性を持っているが、彼らのウィンドウは距離のジャンプが見つかったときのみ作成され、我々はスキャンポイントごとにウィンドウを作成する。
- しかし、最大の違いは、彼らはウィンドウごとに検出中心とオブジェクトクラスを直接予測するのに対し、我々は中心を投票することであり、これは我々の検出器において不可欠であることが示されているステップである。
- さらにいくつかの文献[7]、[14]、[17]では、距離不変表現の必要性が指摘されており、我々はこの研究においてそれを強調するものである。
- しかし、レイトレーシングによって3D点群から再サンプリングされた深度画像を作成する解決策[7]は、我々が提案した単純な前処理よりも著しく複雑な複数の操作を含み、タイミングも提供されていない。
- 他のよく使われる解決策は、3次元占有格子の作成である [14], [17].
- このようなアプローチは、N次元多様体上のデータに対してN+1次元の入力を効果的に用いるが、我々のアプローチでは入力にはN次元のままである。
- 後者の方が学習が容易であり、予測もより頑健であることは言うまでもない[3]。
- さらに、占有格子の「穴埋め」については、複数の異なるヒューリスティックが存在する[17]が、我々の前処理ではそもそも穴が生じないのである。
- 結局のところ、どちらのデータ表現も有効であるが、我々の低次元表現の方がより単純で効果的であると考える。
- 他の人はレーザースキャンで検出するために投票を使用しています。



- Mozosら[20]はマルチハイトレーザースettingsアップで、Spinelloら[29]は3Dレンジデータに基づいたレイヤーファッションで。
- 両者とも、[15]と関連して、データから形状モデルを学習し、異なるレーザ層から検出セントロイドに投票を投じている。
- しかし、Weinrichら[34]と同様に、ジャンプ距離のセグメンテーションに依存し、セグメントのみが検出のために投票することができる。
- Wangらは、[33]において、投票スキームを用いた検出が、線形モデルに対する疎なグリッド上のスライディングウィンドウ検出に相当することを証明している。
- 彼らは、手作業で作成した特徴と線形SVMを用いて点群から良好な検出を実現していますが、学習した非線形特徴や分類器は、手作業や線形分類器に大きな差をつけて勝ることが、コンピュータビジョンの文献で（そして本論文で）何度も示されています。
- また、我々の知見と同様に、最近、投票は他の様々な深層学習アプローチと組み合わせうまく機能することが示されている[19]、[25]、[36]。
- したがって、DROWのような深い非線形投票検出器と[33]で示されるような疎な入力に対するスライディングウィンドウ検出器の関係を確立することは興味深いことであろう。
- レンジデータでの検出は、人物や移動支援機器に限らない。本論文のプレプリント6をアップロードした頃、Ondruskaら[22]は、2次元レンジデータにおける歩行者、自転車、バス、自動車、道路障害物の「追跡」を行う方法を示しています。
- 一見我々の研究と似ていますが、彼らの入力は占有グリッド（つまり $N+1$ 次元）であり、ID付きの離散検出やトラックとは対照的に、ラベル付き占有グリッドを予測しています。
- 彼らのRNNの静的グリッドセルに対するバイアスに基づき、彼らのアプローチが移動ロボットでどの程度機能するかは未知数である。
- Merdrignacら[18]は、2Dレンジデータで自動車、自転車、静的な道路障害物を検出する。
- 彼らは、レンジデータから従来よりも多くの情報を抽出できると仮定し、手作業で作成した大規模な特徴量のセットを設計することによってこれを達成することを目指しています。
- 我々はこれに同意するが、その代わりに、データから直接特徴表現を学習する。



# VI. CONCLUSIONS

2022年6月16日 20:24

- 本論文では、2Dレンジデータから車椅子と歩行者のための高速なディープラーニングベースの検出器であるDROW検出器を紹介しました。
- 我々は深度前処理と投票スキームを提案し、その両方によってCNNがナイーブCNN検出ベースラインを大きく上回り、以前の方法と比較して最先端の結果を得ることができる。
- 我々は徹底的な実験評価を行い、我々の主要な設計上の選択をすべて正当化した。
- また、車椅子と歩行者の中心座標を記録した大規模なデータセットを作成し、今後の研究に役立てたいと考えています。
- 我々は、我々の検出器が、学習データがあれば他のクラスにも一般化することを確信しており、その結果、コミュニティにとって有用であることを確信している。
- 論文が採択された場合、ROSノードと注釈付きデータセットを含む我々のコードを公開する予定です。