教示例の人工的拡張を用いた，深層学習による距離画像からの物体認識

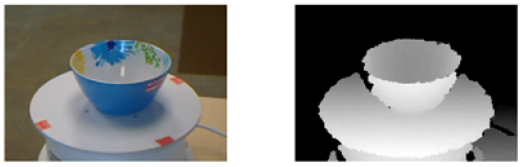
堀内隆志 [1,a)]　古屋貴彦 [1,b)]　大渕竜太郎 [1,c)]

概要：深度センサの普及により，距離画像からの物体認識に対する要求が高まっている．近年，3 次元畳み込みニューラルネットワーク(3DCNN)が距離画像からの物体認識に導入され，一定の効果があることが示されている．しかし，多数のラベル付き距離画像を人手で作成する事が大変な手間であるため，従来の 3DCNN は学習データが不足する傾向にある．本研究は，3DCNN をより効果的に学習して認識精度を向上させることをねらう．ラベル付き 3 次元 CAD モデルからラベル付き距離画像を人工的に多数生成し，これを深度センサで獲得した少数のラベル付き距離画像と併せて 3DCNN の学習に用いる．評価実験の結果，人工的に生成された距離画像の利用により，認識精度が向上することが分かった．

キーワード：深度センサ，距離画像，物体認識，深層学習，3 次元畳み込みニューラルネットワーク

1. はじめに

近年，実世界を認識し行動するロボット (例えば，産業用ロボットや自動運転車など)の発展が目覚ましい．ロボットの自動制御では，色情報を持つ自然画像に加え，距離画像がロボットの「視覚」として利用される．距離画像とは，深度センサを用いて撮影した，センサと物体との間の距離情報を画素値として持つ 2 次元画像である．距離情報は，ロボットが物体を掴むためや自動運転車が衝突を予測するために重要である．図 1 に，RGB-D カメラを用いて，同一の物体を単一視点から撮影して得た自然画像と距離画像の例を示す．図 1 の距離画像において，白い部分は深度センサからの距離が近い場所，黒い部分は深度センサからの距離が遠い場所を表現する．距離画像には，距離センサの視点から見た物体の形状が写る一方で，自然画像に含まれる色情報や物体表面の模様などの情報が存在しない．



自然画像　　距離画像

図 1 自然画像と距離画像の例．
(RGB-D Object Dataset [1])

距離画像中に存在する物体を精度よく認識することは現状，困難である．その理由として，(1) 自然画像と比べて，距離画像には色やテクスチャといった認識のための手掛かりが少ないこと，(2) 深度センサで撮影した距離画像には，一般的に，高周波ノイズや物体の輪郭部に割れ・欠けが含まれること，が挙げられる．さらには，実世界に存在する物体の形状，大きさ，向きが多様であり，それらが互いに重なって隠蔽が生じることで認識がより困難となる．

近年，2 次元畳み込みニューラルネットワーク (2DCNN) が自然画像からの物体認識において高い精度を示している．2DCNN の成功に習い，3 次元畳み込みニューラルネットワーク (3DCNN)を用いた距離画像からの物体認識が研究されている．3DCNN を用いる場合は，物体までの距離情報は 2 次元距離画像ではなく 3D 点群として表現される．図 2 に，深度センサで獲得した 3D 点群の例を示す．図 2 において，撮影視点から見える物体の表面には点群が生成される一方で，撮影視点から見えない物体表面の点群が欠損する．Song ら [3]の手法は，深度センサで獲得した物体の 3D 点群をボクセル表現に変換し，ボクセル表現を 3DCNN へ入力して物体のカテゴリ分類を行う．Song らの手法は，人手で設計された特徴量と「浅い」分類器を組み合わせた従来法よりも高い認識精度を示す．しかし，既存の 3DCNN の認識精度は現状，ロボットの自動制御等の実用には不十分である．認識精度が低い主な理由は，3DCNN の学習に利用可能なラベル付き点群データが少ないためである．一般的に，3DCNN を含む深層ニューラルネットワーク (DNN)を効果的に学習するためには，非常に多数のラベル付きデータが必要である．しかし，人手で多数のラベル付き距離画像を作成するには大変なコストを要する．また，自然画像の場合と異なり，ウェブを利用してラベル付き距離画像を大量に収集することも現実的でない．

本研究の目標は，深度センサで獲得した物体の 3D 点群を高精度にカテゴリ分類することである．3DCNN の学習

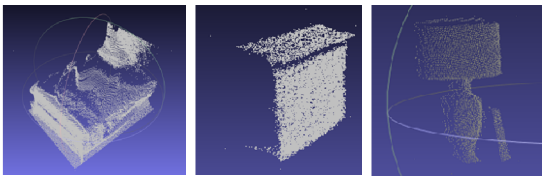


図 2　深度センサで獲得した物体の 3D 点群の例．

1　山梨大学
University of Yamanashi
a)　g16tk013@yamanashi.ac.jp
b)　takahikof@yamanashi.ac.jp
c)　ohbuchi@yamanashi.ac.jp

データの不足を補うために本研究では，ラベル付き 3D 点群データの人工的拡張を行う．具体的には，深度センサで得た 3D 点群よりも多数入手可能な 3D CAD モデルを利用する．3D CAD モデルを擬似的な深度センサで様々な向きから撮影することで, 多様かつ多数の 3D 点群を生成する．データ拡張で得た多数の 3D 点群と，深度センサで得た少数の 3D 点群とを併用して 3DCNN を訓練する．

提案手法の効果を評価するためには，同様の物体カテゴリで構成される 3D 点群データセットと 3D CAD モデルデータセットが必要である．そのようなデータセットは現状存在しないため，既存のデータセットを利用して新たに作成する．評価実験の結果，データ拡張で得た 3D 点群と深度センサで得た 3D 点群の双方を用いて訓練された 3DCNN は，深度センサで得た少数の 3D 点群だけで訓練された 3DCNN よりも高い識別精度を示すことが分かった．

2. 関連研究

距離画像と自然画像を組み合わせた RGB-D 画像から物体を検出し，認識する先行研究として[2][3]などがある．Song らの手法 [2]は，人手で設計した 3D 幾何特徴量を元に「浅い」識別器を学習し，3D の Sliding window 法を用いて 3D 点群シーンからの物体認識を行う．[2]では，まず複数の 3D CAD モデルを多視点から z-buffer レンダリングして 2 次元距離画像を多数作成する. 作成した距離画像を 3D 点群へ変換し，点群から Truncated Signed Distance Function (TSDF)特徴を抽出する．TSDF は，レンダリング視点からの深さ情報に基づく 3D 幾何特徴量である．識別器には Exemplar-SVM [10]が用いられ，物体カテゴリ毎に Exemplar-SVM を学習する．[2]は，3D 点群シーンからの物体認識に 3D の Sliding window を用いる．3D 点群シーンから多様な位置，大きさの 3D 矩形領域を抽出し，3D 矩形領域の各々を Exemplar-SVM を用いて物体であるか否かを判定する．[2]の手法は椅子や机等の比較的単純な形状の 3D 物体の検出に成功するが，3D 点群の欠損やノイズに弱い．この理由は，学習と認識に用いるデータの様態 (モード)が異なるためである．即ち，欠損やノイズを含まない 3D CAD モデルのみを用いて教示した識別器が, 深度センサで獲得した 3D 点群の識別に適するとは限らない．また，人手で設計した特徴量と浅い識別器では，認識能力が限られる可能性がある．

Song らは後に, 深層学習を導入して物体認識精度を改善させた [3]. [3]の手法は, 識別器に 3DCNN を用い, 3DCNN の教示に 3D CAD モデルではなく深度センサで撮影された距離画像を用いる点で[2]の手法と異なる．[3]の手法は，RGB-D センサで得た距離情報と色情報の双方を用いて物体認識を行う．3D 点群から 3D 位置情報付きの TSDF 特徴を抽出し，これを 3DCNN へ入力する．また，色情報を含む自然画像を 2DCNN へ入力する．3DCNN と 2DCNN の識別結果を統合し，物体であるか否か判定する．また，[3]は Sliding window 法の計算量を削減するために，物体の位置を提案する Region Proposal Network (RPN)を利用する．RPN は元々，2 次元画像からの物体検出に用いられる [4]が，Song らはこれを 3D に拡張した．RPN は畳み込み処理により物体位置の推定が可能であるため，3 次元空間の局所領域を 1 つ 1 つ確認する[2]の Sliding window よりも高速な物体認識が可能である．

[3]の手法は[2]の手法より高い認識精度を示したものの，その認識精度はロボットの自動制御等の実用には不十分である．認識精度が低い主な理由は，3DCNN の教示に用いたラベル付き 3D 点群が少ないためと考えられる．[3]の識別器の教示には，SUN RGB-D [5]データセットで用意された 3 次元アノテーション付き距離画像群を用いる．SUN RGB-D には 1 万枚の距離画像に約 6.4 万個分の 3 次元アノテーションが用意されている．しかし，3 次元アノテーションが存在しても隠蔽や欠損により点群が殆ど存在せず，形状を成さない物体が多数存在する．また，6.4 万個の点群データ中に数個から数十個しか存在せず，教示と識別が不可能な物体カテゴリも多く存在する．そのため，物体カテゴリ判別機の教示とテストに利用できるのは[5]のデータセット中において物体 4 万個分である. このデータ数は，3DCNN の教示には不十分であると考えられる．例えば，2DCNN の学習とテストに用いられる CIFAR-10 データセット [11]は 32×32 画素の画像 6 万枚から成る．3D 点群は 2 次元画像よりもデータの次元数が高いため，より多くの(かつ多様な)教示例を要する可能性がある．

3. 提案手法

本研究では，学習データの不足を補い，3DCNN を効果的に学習するために，ラベル付き 3D 点群データの人工的拡張を行う. 一般的に, 深度センサで 3D 形状を獲得する際，物体をどの視点から撮影するかは不定であり，かつ，撮影された 3D 点群には点群の欠損やノイズが含まれる．我々は，これら 3D 点群の撮影条件と性質に合わせたデータ拡張法を提案する．

提案するデータ拡張の手法は 2 つある．1 つ目は，Simulated Range Scan (SRS)法 [6]を用いたデータ拡張(*data Augmentation by Simulated range scan, AS*)である (3.1 節)．AS 法は，ラベル付き 3D CAD モデルを擬似的な深度センサで様々な向きから撮影することで，点群の欠損やノイズを含む多様かつ多数の 3D 点群データを生成する．2 つ目は，3D モデルの回転によるデータ拡張 (*data Augmentation by Rotation, AR*)である (3.2 節)．AR 法は，3D CAD モデルを物体の上向きベクトルを中心として回転することで，多様な向きの，かつ，欠損のない 3D 形状デー

タを生成する．

3.1 3D 点群データの人工的拡張

3.1.1 Simulated Range Scan を用いたデータ拡張 (AS 法)

AS 法はまず，ラベル付き 3D CAD を多視点から深さ値レンダリングする．図 3 に，AS 法によるデータ拡張の概要を示す．3D CAD モデルはその上向き方向が揃っており，かつ，実世界の物体は底面から撮影されることはない仮定に基づき，次の方法でレンダリングする．3D CAD モデルを図 3 に示す 13 面体で囲み，13 面体の底面を除く床上 12 個の面の重心に撮影視点を置き，3D CAD モデルの重心に視線を向けて z-buffer レンダリングする．

レンダリングして得た深さ画像の各画素を，SRS 法を用いて，3D 点群へ変換する．図 4 に，本手法で生成された 3D 点群の例を示す．深度センサによる撮影と同様，SRS 法で生成される 3D 点群は 3D CAD モデルの自己隠蔽により，撮影視点から見えない部位の点群が欠損する．3D 点群のラベルは，生成元の 3D CAD モデルのラベルと等しい．

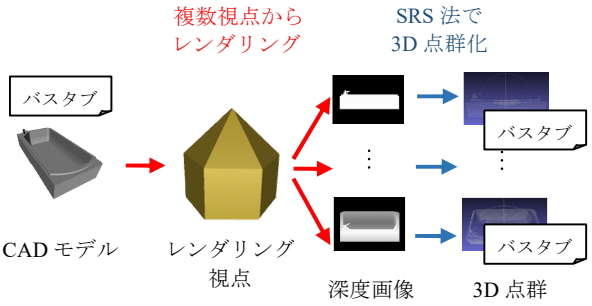


図 3 ラベル付き 3D CAD モデルから複数視点のラベル付き 3D 点群を作成することで教示例を拡張．

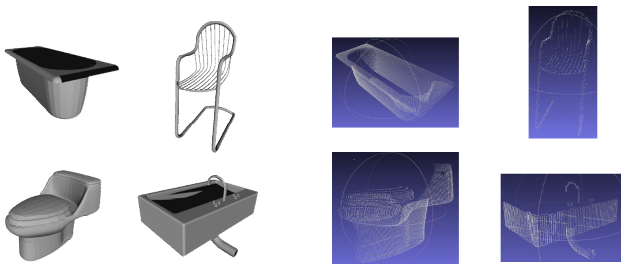


図 4 3D CAD モデルと，AS 法で生成された 3D 点群の例．

また本研究では，点群の欠損部位の存在を考慮した 3D 点群を生成することで，認識精度の改善を試みる．このために，Amodal bounding box [3][6]を利用する．具体的には，変換後の 3D 点群に，変換前の 3D CAD モデルの頂点座標の x，y，z 値の最大値，最小値を取る点 8 個を追加する．図 5 に，Amodal bounding box の例を示す．Amodal bounding box の利用により，距離情報に加え，物体全体に対しどの程度の割合が撮影されたのかを認識の手掛かりとすることができる．

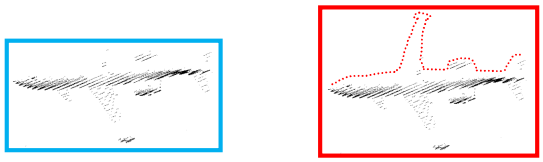


図 5 点群を囲む bounding box と点群の欠損部位を考慮した amodal bounding box．提案手法では，amodal bounding box を 3DCNN の教示に用いる．

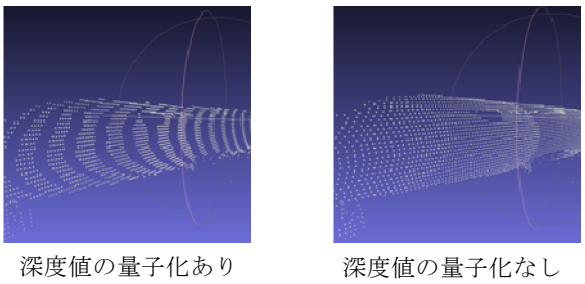


図 6 深度値の量子化の有無の違い．提案手法では，量子化ありの点群を 3DCNN の教示に用いる．

一般的に，深度センサで得られる距離画像はその画素値が例えば 256 段階に量子化される．また，深度センサ毎の解像度の違いから，距離画像を変換して得た 3D 点群が持つ点数も深度センサ毎に異なる．そこで本研究では，深度センサで得る 3D 点群により類似した 3D 点群を生成するために，深度値の量子化，および，点の数の制御を導入する．図 6 に，深度値の量子化の有無による 3D 点群の違いを示す．以下では，深度センサの機種毎の特性を考慮する場合としない場合の 2 種類のデータ拡張法を説明する．

センサ機種の特性の考慮なし：3D CAD モデルを 500×500 画素の解像度で 12 視点から z-buffer レンダリングし，3D 点群を生成する． 深さ画像の背景 (無限遠)を除く画素群の内，最も深度値の小さい (レンダリング視点に近い)部分から，最も深度値の大きい (レンダリング視点から遠い)部分までを 300 段階に均等に分割し，各画素の深度値を，最も値の近い段階に量子化することで，点群の物体表面に量子化誤差を付与する．

センサ機種の特性の考慮あり：本論文の実験で用いる SUN RGB-D は，Microsoft 社が販売する Kinect V1, Kinect V2, Asus 社の Xtion，Intel 社の Real Sense の 4 種類の深度センサで撮影された RGB-D 画像群から成る．なお，Kinect V1 と Xtion は PrimeSense 社が ODM 生産を 2 社から請け負っており，深度センサの仕様は同一であるため本論文では両センサを合わせて Kinect V1 として扱う．本拡張方法では

まず，各センサで撮影された物体の，点の数と深度値の量子化の段階数を調査する．その後，全てのセンサにおける物体カテゴリごとの点の数と深度値の量子化段階数の平均値求める．表 1 は求めた深度センサごとの物体の点の数，表 2 は深度値の量子化段階数である．

表 1 センサ毎，カテゴリ毎の点群の点の数

カテゴリ名	Kinect V2	Kinect V1	Real Sense	センサ平均
bathtub	64,355	42,276	42,276	49,636
bed	233	55,150	129,303	61,562
bookshelf	41,062	44,271	66,626	50,653
chair	14,503	12,435	38,420	21,786
desk	30,947	32,581	76,109	46,546
door	40,401	22,327	83,870	48,866
lamp	5,306	6,320	14,158	8,595
monitor	9,139	9,727	7,473	8,780
night_stand	11,218	11,671	11,674	11,521
sink	21,554	23,480	28,372	24,469
sofa	58,663	40,681	99,598	66,314
table	29,511	23,885	61,433	38,276
toilet	39,636	25,238	33,145	32,673

表 2 センサ毎，カテゴリ毎の深度値の量子化段階数

カテゴリ名	Kinect V2	Kinect V1	Real Sense	センサ平均
bathtub	121	100	100	107
bed	233	164	174	190
bookshelf	114	85	113	104
chair	70	63	76	70
desk	120	112	148	127
door	53	50	63	55
lamp	47	40	50	46
monitor	48	40	45	44
night_stand	64	62	62	63
sink	65	63	67	65
sofa	167	137	154	153
table	131	108	132	124
toilet	73	59	66	66

カテゴリラベル付き 3D CAD モデルを z-buffer レンダリングする際には，表 1 および表 2 に記載された平均の点数と平均の量子化段階数に従って 3D 点群を生成する．事前実験では平均ではなくセンサ機種ごとの点数，量子化段階数に従い 3D 点群を作成し，これらを混ぜて 3DCNN の教示に用いたが識別精度が低下した．センサ機種ごとに異なる傾向を持つ 3D 点群が作成されたために識別器の学習が阻害された可能性がある．

3D 点群の点数を表 1 の平均値に近づける方法は次の通りである．まず 500×500 の解像度で 3D CAD モデルを z-buffer レンダリングする．得られた 3D 点群の点の数をn_cとする．表 1 で示すカテゴリごとの各センサの平均の点の数をn_aとする．レンダリング解像度を 1 辺あたり$\sqrt{n_a/n_c}$倍にして再度 z-buffer レンダリングすることで，3D 点群の点の数を各センサの点の数の平均に近付ける．

深度値の量子化段階数を表 2 の平均値に近づける方法は次の通りである．点数を平均化した 3D 点群を得た後，その点群中の最も深度値の小さい点と最も深度値の大きい点を取得する．最小深度点と最大深度点を結ぶ線分を表 2 で示すカテゴリごとの平均量子化段階数で均等に分割し，各点の深度値を，最も値の近い段階値に量子化する．

3.1.2 回転によるデータ拡張 (AR 法)

AR 法は，3D CAD モデルの回転によりデータの多様性を拡張する．図 7 に，AR 法によるデータ拡張の概要を示す．物体の上向き方向を軸に 3D 形状を回転させる．回転させる方向数は，4, 8, または 12 とし，回転の間隔は 360° を方向数で割った値に等しい．

実験では，3D CAD モデル，または，深度センサで得た 3D 点群に対して AR 法を適用する．注意点として，3D CAD モデルから AR 法で生成した 3D 形状は，SRS 法による点群化は行われず，形状全体の情報を保持したまま後述の 3DCNN へ入力される．

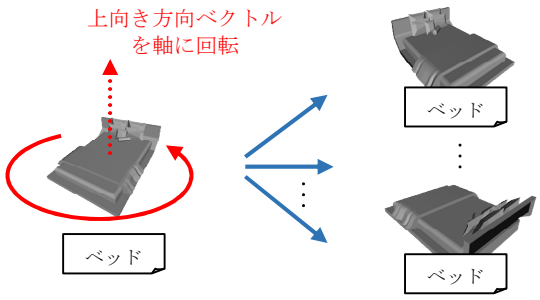


図 7 3D 形状の上向き方向を軸として複数方向に回転することで教示例を拡張．

3.2 3DCNN を用いた 3D 点群の識別

3.2.1 3DCNN の構造

3.1 節で生成したラベル付き 3D 点群データを用いて 3D CNN を学習し，学習後の 3DCNN を認識に用いる．本節では実験に用いる 3DCNN の構造について述べる．なお，本研究で用いる 3DCNN は, state-of-the-art な 2DCNN (例えば, [13][14])と比べて浅い．浅い 3DCNN を用いた理由は，学習の時間を削減し，様々な実験条件を試行するためである．

3DCNN への入力：ボクセルデータを 3DCNN へ入力する．ただし，高解像度なボクセルデータを 3DCNN に入力する

と 3D 畳み込みの計算量が爆発し，現実的な処理時間で学習と識別を行うことが出来ない．そこで，本研究では，既存の 3DCNN [7]と同様，3D 点群を低解像度 (30×30×30) のボクセルへ変換し，これを 3DCNN へ入力する．3D 点群をボクセル化する際は，3D 点群を 30×30×30 の立方体格子で囲み, 各格子に点が 8 個以上在る場合は格子の値を“1”に，そうでない場合は“0”とする．一方で，3D CAD モデルをボクセル化する際は，3D CAD モデルを 30×30×30 の立方体格子で囲み，各格子にポリゴンが含まれる場合は格子の値を“1”に，そうでない場合は“0”とする．

3DCNN の構成 : 本研究では 3D ShapeNet [7]を参考に，図 8 に示す構造の 3DCNN を用いる．この 3DCNN は入力の後に 2 つの 3D 畳み込み層, その後に 2 つの全結合層を持つ．3D 畳み込み層と全結合層の活性化関数に ReLU [12]を用いる．最後の全結合層の次に softmax 層を置くことで，入力されたボクセルデータの物体カテゴリを識別する．

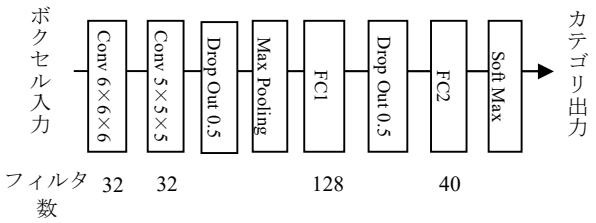


図 8 実験に使用した 3DCNN の構造.

3.2.2 3DCNN の学習

3DCNN の学習の目的は，3DCNN が入力されたボクセルデータの物体カテゴリを正しく識別できるよう，3DCNN のパラメタ (互いに隣接する層を繋ぐ接続重み)を調整することである．3DCNN の教示は 2 段階で行われる．1 段階目は，3.1 節で述べた手法を用いて得た疑似点群データを用いて，物体のカテゴリ識別を出来るように 3DCNN を訓練する (「事前学習」). 事前学習の教示データには，表 2 で示した 13 カテゴリのいずれかに属するラベル付き 3D CAD モデルまたはラベル付き 3D 点群から作成したボクセルデータを用いる. 事前学習の後, 2 段階目の学習として，深度センサで獲得した実点群データを用いて 3DCNN を追加学習する (「微調整」).

3DCNN のパラメタ初期値はランダムに決定する. 事前学習と微調整の損失関数には Multinomial Logistic Loss を用い，逆誤差伝播法を用いて学習する．ソルバーには確率的勾配降下法を用い，ミニバッチのサイズを 32 とする．初期学習係数は 0.001 とし, 4 万バッチの学習ごとに学習係数を 1/10 に減らしていく．事前学習，微調整共に 50 万バッチの学習を行う．

4. 3D 点群データセットの作成

本論文の実験では，深度センサで撮影された距離画像のデータセットとして SUN RGB-D を用いる. SUN RGB-D は 1 万枚の RGB-D 画像に 6.4 万の 3D アノテーションが割り振られており，計 6.4 万個のラベル付き点群を切り出すことが出来る．しかし，そのまま点群を切り出すだけでは以下に示す I〜III の問題が生じる．

I. ラベルのカテゴリ名に表記揺れが存在する問題.
II. 点が少な過ぎて, 物体の形状を成さない点群が存在する問題.
III. 点群をアノテーションに従い切り出しただけでは, 物体の本来の大きさが分からなくなる問題.

I の問題により，事前学習に用いる 3D CAD モデルと微調整に用いる 3D 点群のカテゴリ分布が異なると，3DCNN の学習を阻害する可能性がある．カテゴリ名の表記揺れ問題を解決するべく，人手で表記揺れのあるカテゴリ名を探し出し，統合する作業を行った．表 3 に，統合前と統合後のカテゴリ名を示す.

表 3 SUN RGB-D のカテゴリ名の統合

統合後	統合前
batutub	batutub
bed	bed, sofa_bed, bunk_bed, folding_bed, child_bed, baby_bed
bookshelf	bookshelf
chair	chair, sofa_chair, stack_of_chairs, chairs, lounge_chair, baby_chair, sauceer_chair, child_chair, high_chair, massage_chair, lawn_chair
desk	desk
door	door, lift_door, hingedoor
lamp	lamp
monitor	monitor
night_stand	night_stand, nightstand
sink	sink, kitchen_sink
sofa	sofa
table	table, coffe_table, coffetable, endtable, end_table, dining_table, side_table, changing_table, centertable, laboratory_table, long_office_table, bar_table
toilet	toilet

II の問題への対処として，点の数が 500 未満の点群は距離画像から切り抜かないこととする．III の問題への対処として，3.1 節で述べた Amodal bounding box を利用する．切り出した点群にアノテーションで示された物体の大きさを保存できるよう，物体の x，y，z 座標の最大値，最小値を取る点 8 個を点群に追加する.

上記の操作の結果，SUN RGB-D データセットから，13 個の物体カテゴリに分類された 41,692 個の 3D 点群データが生成される．しかし，“chair”, “desk”, “table”カテゴリのみで 28,590 個のモデルが存在し，カテゴリごとのモデル数のばらつきが大きい．学習・評価へのモデル数の偏りの影響

を減らすべく，モデル数の多いカテゴリからはモデルを間引く．間引いた後の 12,566 個のモデルを実験に使用する．我々はこれを 11,306 個の学習用データと，1,260 個の評価用データに分割して実験に用いる．

5. 実験と結果

5.1 実験条件

提案手法である教示例の人工拡張の効果を確かめるため，距離画像から切り出した 3D 点群のカテゴリ判別の精度を評価する．実験には，4 章で作成した 3D 点群データセットと，40 個の物体カテゴリに分類された 3D CAD データから成る ModelNet40 [7]を用いる．

3DCNN の学習：ModelNet40 の学習用セットから，表 1 に示したカテゴリのラベル付き 3D モデルを計 56,688 個取り出し，教示例の拡張に用いる．教示例の拡張には AS 法と AR 法を用いる．AS 法で拡張する場合，56,688 個の 3D CAD モデルの各々を SRS 法で点群化し，計 421,872 個のラベル付き点群データセットを作成する．実験では，センサ機種毎の特性の考慮の有無 (3.1 節)による精度比較を行う．

一方で，AR 法で拡張する場合，56,688 個の 3D CAD モデルの各々に回転を加えてデータ拡張する．回転方向数は，0 方向 (データ拡張無し)，4 方向，8 方向，12 方向のいずれかである． また，SUN RGB-D から取り出した 11,306 個の教示用の 3D 点群にも，0 方向，4 方向，8 方向，12 方向の 4 種類のデータ拡張を行う．

識別精度評価：精度評価には，SUN RGB-D から取り出した 1,260 個の評価用データを用いる．なお，精度評価時には 3D 点群に対して回転拡張を実施せず，3D 点群をそのままボクセル化して 3DCNN へ入力する．教示例，評価用データのボクセル化解像度は特に言及が無い場合は 30×30×30 とする．評価指標にはカテゴリ正解率を用いる．評価用データのデータ数を N_t，評価用データの内，正しくカテゴリ判別されたデータ数を N_c とするとき，カテゴリ正解率を N_t / N_c と定義する．3DCNN はランダムな初期値やデータを教示する順序等の影響により，学習結果が毎回異なる．そのため各々の実験は 3 回行い，それぞれの結果の平均を報告する．3DCNN の実装と学習，評価には深層学習のツールである marvin [9]を用いた．

5.2 実験結果

5.2.1 事前学習のみを実施した場合のカテゴリ正解率

図 9 に，AR 法で回転拡張した深度センサ 3D 点群，又は 3D CAD モデルを用い事前学習のみ行った 3DCNN の識別精度を示す．深度センサから得られる点群を回転拡張すると識別精度の低下が見られた．深度センサで得た 3D 点群は撮影視点以外の点群が大きく欠損しており，これを回転した 3D 点群を学習した 3DCNN が，深度センサで獲得した (回転のかかっていない)3D 点群の識別に失敗した可能性がある．一方で，3D CAD モデルを回転拡張した際は，回転方向数の増加に伴って識別精度が向上した．3DCNN が物体の多様な向きを学習した為だと考えられる．

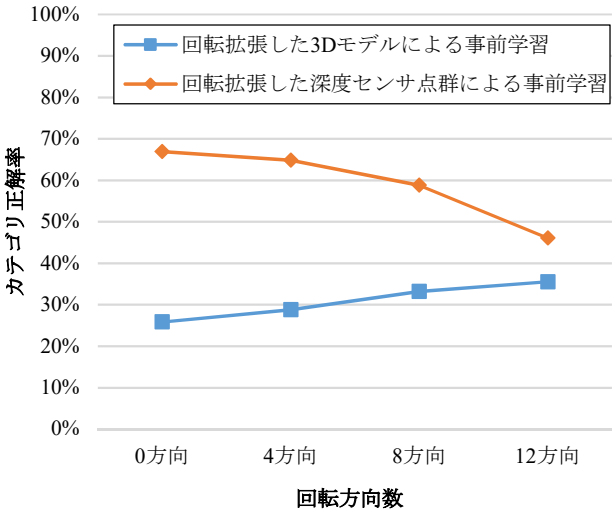


図 9 教示例回転拡張の回転方向数と識別精度の関係

表 4 に，AS 法で作成した 3D 点群で 3DCNN を教示した際の識別精度を示す. 表 4 より, 事前学習を実施する場合，センサ機種の特性を考慮せず 3D 点群を拡張する手法が最も識別精度が高くなることが分かった．

センサ機種特性を考慮すると精度が低下する理由として以下が考えられる．本手法は，物体カテゴリ毎に算出した平均点数と平均量子化段階数に基づいて疑似 3D 点群を生成する. しかしながら, 3D 点群の点数や量子化段階数は，深度センサから撮影対象物体までの距離や撮影時間によりばらつくのが一般的である．点数と量子化段階数のばらつきを考慮しなかった結果，疑似 3D 点群を深度センサで実際に得た 3D 点群に近づけきれなかったと推察する．

表 4 AS 法で拡張した 3D 点群による 3DCNN 識別精度

事前学習に用いた教示例	カテゴリ正解率 [%]
AS 法 3D 点群 センサ特性考慮なし	**41.38**
AS 法 3D 点群 センサ特性考慮あり	35.96
AR 法 3D CAD モデル 12 方向回転拡張	35.55

5.2.2 事前学習と微調整を実施した場合のカテゴリ正解率

表 5 に，3DCNN を AS 法で拡張した 3D 点群，または AR 法で回転拡張した 3D モデルで事前学習し，回転拡張を行っていない深度センサの 3D 点群で微調整を行った際の識別精度を示す．人工拡張した教示例で事前学習を行い，深度センサの 3D 点群で微調整する手法は，深度センサの 3D 点群のみで学習する手法よりも識別精度が高くなった．

また, AS 法 (センサ特性考慮なし)で作成した 3D 点群で事前学習し，深度センサの 3D 点群で微調整を行う手法が，他の学習方法よりも識別精度が若干高くなった．

表 5 事前学習と微調整を実施した 3DCNN 識別精度

事前学習	微調整	カテゴリ正解率 [%]
AS 法 3D 点群 (センサ特性考慮なし)	深度センサ 3D 点群 (回転拡張なし)	**70.42**
AS 法 3D 点群 (センサ特性考慮あり)	深度センサ 3D 点群 (回転拡張なし)	68.36
AR 法 3D CAD モデル (12 方向)	深度センサ 3D 点群 (回転拡張なし)	69.66
なし	深度センサ 3D 点群 (回転拡張なし)	66.95

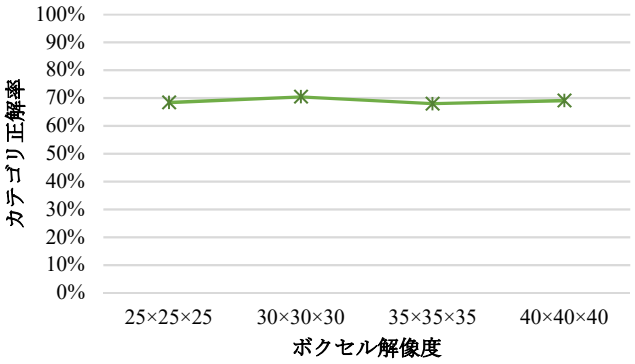


図 10 ボクセル解像度と 3DCNN の識別精度の関係

5.2.3 ボクセル解像度によるカテゴリ正解率

図 10 に, 3D CNN の教示例と評価用データのボクセル解像度を変えた際の識別精度の違いを示す. 本節の実験では，AS 法でセンサ特性を考慮せず拡張した 3D 点群で 3DCNN を事前学習し，深度センサの 3D 点群で微調整を行う．学習に用いる教示例のボクセル解像度を $n \times n \times n$ (n = 25, 30, 35，40)にした際の識別精度を調査した．図 10 より，n=30 の時にカテゴリ正解率が最高の 70.42%，n=35 の時に最低の 68.00%を示すが，ボクセル解像度の違いによる識別精度の大きな変化は見て取れない．

6. まとめと今後の課題

本研究では，距離画像からの物体認識の精度を改善するため，3D 形状データの人工的拡張を行い，3DCNN の学習用データの不足を補った．3D CAD モデルを Simulated Range Scan 法に基づく疑似的な深度センサでスキャンすることで，実際の深度センサで得る 3D 点群に類似した疑似 3D 点群を多数生成した．また，3D CAD モデルを多方向に回転させることでも，教示データの多様性を拡張させた．

新たに作成した 3D 点群データセットを用いた評価実験の結果，教示データの人工的拡張は識別精度の改善に有効であることが分かった．3DCNN を，3D CAD モデルを元に人工的拡張した 3D 点群で事前学習することで，深度センサから得られる物体の 3D 点群のみで教示する場合と比較して，カテゴリ正解率が向上することを確認した．

今後の課題は，さらなる高精度化である．提案手法のカテゴリ正解率は未だ 7 割程度であり実用には不十分である．提案手法の 3DCNN への入力は，“0”または“1”の 2 値を取るボクセル表現である．物体の形状の詳細をより捉えられる 3D 幾何特徴量の導入が考えられる．また，3DCNN は，点群に発生する欠損部分にも畳み込み処理を行い，欠損部分からも特徴を抽出する．欠損部分からは特徴を抽出しない，あるいは欠損部分は物体カテゴリの判断に用いない等の工夫を検討する．

また，本研究ではあらかじめ距離画像中から切り出された物の点群のカテゴリの判断を行っていた．今後は，室内や屋外の 3D 点群シーンから物体の位置を検出する物体検出器の開発も行い，本研究で提案した識別器との連結を実施したい．

参考文献

[1]“RGB-D Object Dataset”. https://rgbd-dataset.cs.washington.edu/.(参照 2017-3-30).
[2] S.Song, and J Xiao. Sliding Shapes for 3D Object Detection in Depth Images. *ECCV2014*, 2014.
[3] S.Song, and J Xiao. Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images. *CVPR2016*, 2016.
[4] S Ren, K He, R Girshick, and J Sun. Raster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *NIPS*, 2015.
[5] S.Song, S. Lichtenberg, and J.Xiao. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. *CVPR2015*, 2015.
[6] Sipiran, I, Meruane. R, Bustos,B, Schreck. T, Li,B. Lu,Y, and Johan, H. A benchmark o simulated range images for partial shape retrival, *The Visual Computer30 (2014), No.11*, pp.1293-1308, 2014.
[7] Z. Wu, S. Song , A. Khosla, F.Yu, L. Zhang, X. Tang, and J.Xiao.3D ShapeNets: A Deep Representation for Volumetric Shapes.*CVPR2015*, 2015.
[8] “The Princeton ModelNet.”. http://modelnet.cs.princeton.edu/. (参照 2017-3-30).
[9] “Marvin A minimalist GPU-only N-dimensional ConvNet framework”. http://marvin.is/.(参照 2017-3-30).
[10]T. Malisiewicz, A. Gupta, and A. Efros. Ensemble of Exemplar-SVMs for Object Detection and Beyond.*ICCV11*, 2011.
[11] “CIFAR-10 and CIFAR-100 datasets”, http://www.cs.toronto.edu/~kriz/cifar.html. (参照 2017-4-10).
[12] X. Glorot, A. Bordes, and Y. Bengio. Deep Sparse Rectifier Neural Networks. Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTAS-11), 2011.
[13] K. He, X. Zhang, S. Ren, and J. Sun, Deep Residual Learning for Image Recognition, *CVPR 2016*, 2016.
[14] C. Szegedy et al. Going Deeper with Convolutions, *CVPR 2015*, 2015.