1C-01

立体フィルタを用いた畳み込みニューラルネットワークによる三次元物体認識

金井 廉† 藤田 悟†

法政大学大学院 情報科学研究科 情報科学専攻†

1 はじめに

近年，Kinect や 3D スキャナ等の三次元デバイスが普及し，膨大な量の三次元データが蓄積され続けている．また，それに伴ったアプリケーションの普及とともに，三次元物体認識の必要性も高まりつつある．一方で，深層学習というニューラルネットワークを多層構造にした識別器を大量のデータで学習させる手法が近年盛んに研究されており，画像や音声等を高精度で認識できることが報告されている．本論文では，深層学習で特に画像認識に用いられる畳み込みニューラルネット（Convolutional Neural Network，以下 CNN）[1]を応用した三次元物体認識手法を提案する．CNN は画像中の二次元的な特徴を抽出するが，本手法では CNN を三次元に拡張し，三次元形状の形状特徴を抽出することで高精度な三次元物体認識を実現する．

2 関連研究

2.1 CNN

CNN は畳み込み層，プーリング層と呼ばれる層を持つニューラルネットワークである．畳み込み層は，入力画像に対しフィルタと呼ばれる二次元の重み信号を用いた畳み込み演算を行う層であり，フィルタの表す特徴的な濃淡構造を入力画像から抽出する．入力画像のチャネル数をC，サイズを$W\times W$画素，フィルタの数をM，サイズを$H\times H$画素とし，入力画像中の画素(i,j)（$1\le i,j\le W-1$）のチャネルc（$1\le c\le C-1$）の値をx_{ijc}，チャネルcの画素値に対するm（$1\le m\le M-1$）番目のフィルタ中の画素(p,q)（$1\le p,q\le H-1$）の値をh_{pqcm}とすると，フィルタ適用位置の移動間隔（ストライドという）をs画素としたときの CNN の畳み込み演算は以下のように定義される．

$$u_{ijm}=\sum_{c=0}^{C-1}\sum_{p=0}^{H-1}\sum_{q=0}^{H-1}x_{si+p,sj+q,c}h_{pqcm} \qquad (1)$$

畳み込み演算結果は二次元画像となり，画素値u_{ijm}は各畳み込み領域のパターンに対するフィルタの反応度合いを表す．なお，畳み込み層の最終的な出力は，活性化関数による写像となる．プーリング層ではこの出力を入力画像とし，画像中の特定領域から一つの画素値を出力する．出力方法はいくつか存在するが，一般に最大プーリングがよく用いられる．最大プーリングは，サイズ$W\times W\times C$の入力画像に(i,j)を中心としたサイズ$H\times H$の領域をとり，その中の画素集合の最大値を出力とする．プーリング層は畳み込み層でのフィルタの反応の位置ずれを吸収するので，平行移動に頑強な識別が可能となる．プーリング層の出力も畳み込み層同様二次元となり，入力画像から抽出された特徴量として全結合のニューラルネットワークや他の識別器の入力となる．

2.2 CNN を用いた三次元物体認識

RGB-D の画像を学習データとする CNN の三次元物体認識の例が既存研究として報告されている[2]が，RGB-D の画像は深度カメラ等で取得しやすい反面，一視点からの色及び深度情報のため，本来物体が持つ三次元的な形状情報が失われやすい．本手法では三次元形状を直接 CNN の入力とすることで，対象データの三次元的な情報を失わずに認識を行うことを目指している．三次元形状データを得られる状況であれば，高い精度で認識を行うことができる．

3 提案手法

本論文では，三次元形状データをボクセルに変換して入力とし，三次元の立体フィルタによるボクセル単位の畳み込み，プーリングを行う CNN（以下三次元 CNN）を提案する．入力データのチャネル数をC，サイズを$W\times W\times W$，フィルタの数をM，サイズを$H\times H\times H$とし，入力デー

3D Object Recognition using Convolutional Neural Networks with 3D Filters

† Ren Kanai, Satoru Fujita

† Graduate School of C.I.S., Hosei University

表 1 三次元 CNN の構成

layer	$H\times H\times H$	s	output	$f(x)$
input	-	-	100×100×100×1	-
conv	5×5×5	3	32×32×32×16	ReLU[4]
pool	4×4×4	4	8×8×8×16	-
fc	-	-	1×1×1×512	ReLU
fc	-	-	1×1×1×256	ReLU
fc	-	-	1×1×1×5	softmax

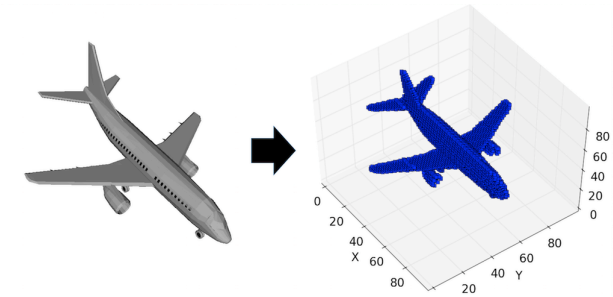

図 1 三次元形状のボクセル変換例

タのボクセル (i,j,k) $(1\le k\le W-1)$ のチャネルcの値をx_{ijkc}，チャネルcのボクセル値に対するm番目のフィルタ中の画素 (p,q,r) $(1\le r\le H-1)$ の値をh_{pqrcm}とすると，三次元 CNN の畳み込み演算は以下の式で表される．

$$u_{ijkm}=\sum_{c=0}^{C-1}\sum_{p=0}^{H-1}\sum_{q=0}^{H-1}\sum_{r=0}^{H-1}x_{si+p,sj+q,sk+r,c}h_{pqrcm} \qquad (2)$$

また本手法におけるプーリング層では最大プーリングが行われ，ボクセル (i,j,k) を中心としたサイズ$H\times H\times H$の領域中の最大値を返す．畳み込み層で三次元空間上の形状パターンに対するフィルタの反応が出力され，プーリング層で平行方向の反応のずれが吸収される．本手法ではこの畳み込み層とプーリング層によって抽出された特徴量を全結合の層に入力する．

4 実験

本手法の認識精度を検証するため，三次元物体認識の評価に用いられる三次元形状データセット Princeton Shape Benchmark（PSB）[3]データセットを用いた実験を行った．今回の実験では 5 クラスのいずれかに属する訓練データ，テストデータを共に 102 個用意し，各形状データをモデルの重心のz軸まわりに 45 度まで 5 度ずつ回転させ，918 個とした．各形状データは図 1 に示すような分割数100×100×100のボクセルへの変換がなされ，三次元 CNN に入力される．本手法で用いる三次元 CNN の構成を表 1 に示す．

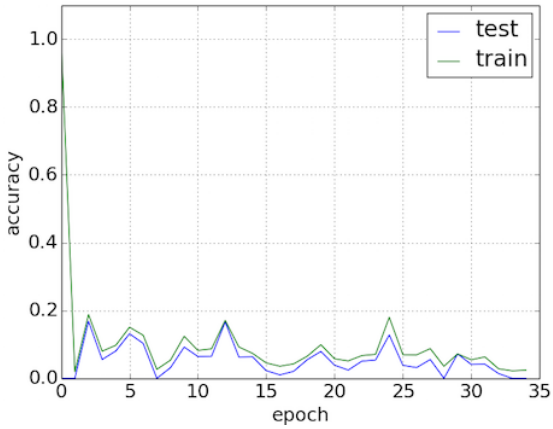

図 2 誤り率の遷移

5 実験結果

訓練データ及びテストデータの誤り率の遷移をエポックごとにプロットしたグラフを図 2 に示す．エポックごとに精度の振動が見られるが，いずれも誤り率は概ね 20%以下に収まっていることがわかる．

6 まとめ

本論文では，高精度な三次元物体認識を行うため，従来手法である CNN を三次元に拡張する手法を提案した．三次元 CNN ではボクセル値のない部分への畳み込みも行うため，今後はそのような無駄のないニューラルネットを構築する必要があると考える．

参考文献

[1] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, “Gradient-based learning applied to document recognition,” In Proceedings of the IEEE, 86, pp.2278-2324, 1998.

[2] R. Socher, B. Huval, B. Bhat, C. D. Manning and A. Y. Ng, “Convolutional-Recursive Deep Learning for 3D Object Classification,” In Advances in Neural Information Processing Systems 25, 2012.

[3] P. Shilane, P. Min, M. Kazhdan and T. Funkhouser, "The Princeton Shape Benchmark, " Proc. Int'l Conf. On Shape Modeling and Applications 2004(SMI '04), pp.167-178, 2004.

[4] Glorot, Xavier, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks,” International Conference on Artificial Intelligence and Statistics, pp.315-323, 2011.