一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
PRMU2016-39,SP2016-5,WIT2016-5(2016-06)

深層学習を用いた多様体構築による 3次元物体の姿勢推定に関する予備検討

二宮　宏史† 　川西　康友† 　出口　大輔†† 　井手　一郎† 　村瀬　　洋†

小堀　訓成††† 　橋本　国松†††

† 名古屋大学 大学院情報科学研究科 〒 464–8601 愛知県名古屋市千種区不老町
†† 名古屋大学 情報戦略室 〒 464–8601 愛知県名古屋市千種区不老町
††† トヨタ自動車株式会社 〒 471–8571 愛知県豊田市トヨタ町 1
E-mail: †ninomiyah@murase.m.is.nagoya-u.ac.jp, ††ddeguchi@nagoya-u.jp,
†{kawanishi,ide,murase}@is.nagoya-u.ac.jp, †††{norimasa_kobori,kunimatsu_hashimoto}@mail.toyota.co.jp

あらまし　近年，物体の 3 次元姿勢推定技術が注目されている．従来技術として，3 次元姿勢変化による 2 次元画像上での見えの変化を，低次元空間中における多様体で表現するパラメトリック固有空間法がある．この手法は見えの分散に着目し，主成分分析を用いて特徴量を求めるものである．しかし，姿勢変化の大きさと見えの変化の大きさには必ずしも関係があるわけではない．そのため，見えの変化が小さい姿勢の違いを区別することが難しいという問題がある．本報告では，姿勢の分離性に着目した特徴量による多様体構築手法を提案する．姿勢を教師信号として学習した Deep Convolutional Neural Network（DCNN）の中間層から，姿勢の分離性が高い特徴を抽出する．この特徴量を用いて多様体を構築することで，見えの変化が小さい姿勢の違いも区別できると考えられる．実験により，姿勢を教師信号として学習した DCNN から抽出した特徴量による多様体構築を行なうことで，姿勢推定精度の向上を確認した．
キーワード　3 次元物体，姿勢推定，多様体，深層学習

Preliminary study on deep manifold embedding for 3D object pose estimation

Hiroshi NINOMIYA†, Yasutomo KAWANISHI†, Daisuke DEGUCHI††,

Ichiro IDE†, Hiroshi MURASE†, Norimasa KOBORI†††, and Kunimatsu HASHIMOTO†††

† Graduate School of Informaion Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, 464–8601 Japan
†† Information Strategy Office, Nagoya University
Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, 464–8601 Japan
††† Toyota Motor Corpolation
1 Toyota-cho, Toyota-shi, Aichi, 471–8571 Japan
E-mail: †ninomiyah@murase.m.is.nagoya-u.ac.jp, ††ddeguchi@nagoya-u.jp,
†{kawanishi,ide,murase}@is.nagoya-u.ac.jp, †††{norimasa_kobori,kunimatsu_hashimoto}@mail.toyota.co.jp

Abstract　Recently, 3D object pose estimation is being focused. The parametric eigenspace method is known as one of the fundamental methods. It represents the appearance change of an object caused by pose change with a manifold embedded in a low-dimentional subspace. It obtains features by PCA, which maximizes the appearance variation. However, there is not always a correlation between pose change and appearance change. So, there is a problem that the method cannot handle a pose change with a slight appearance change. In this report, we introduce deep manifold embedding which maximizes the pose variation. We construct a manifold from features extracted from Deep Convolutional Neural Networks (DCNNs) trained with pose information. Pose estimation with the proposed method achieved the best accuracy in experiments using a public dataset.
Key words　3D object, pose estimation, manifold, deep learning

This article is a technical report without peer review, and its polished and/or extended version may be published elsewhere.
Copyright ©2016　by　IEICE

1. まえがき

ロボットの導入が産業分野や生活分野等に進みつつある．産業分野において，生産現場での部品のピッキングがロボットの導入によって自動化されている．近年ではこの技術の物流分野への展開を目指し，実環境における自動ピッキング装置の開発を目的とした Amazon Picking Challenge [1] というコンペティションが開催された．また，生活分野においても，少子高齢化の進展に伴う介護・福祉，家事等の労働力不足の懸念から，生活支援を目的としたロボットが開発されている [2]．両分野とも，ロボットが物体を掴むことが共通の課題であり，物体を掴むための技術が求められている．ロボットが物体を掴むには，対象物体を掴める方向を知るために，物体の 3 次元姿勢推定を行なう必要がある．

また，ロボットに搭載するセンサとして 3 次元ビジョンセンサなどが考えられるが，価格が高い，サイズが大きい，使用環境やデータ取得可能な対象物体が限定的，キャリブレーションが煩雑など，実用上の課題も多い．このため，3 次元ビジョンセンサではなく従来の単眼カメラからの物体の 3 次元姿勢推定技術が求められている．

単眼カメラからの物体の 3 次元姿勢推定手法として，回帰モデルに基づく手法 [3] とテンプレートマッチングに基づく手法 [4] がある．前者では，あらかじめ画像から得られる特徴量とそれに対応する姿勢を獲得しておき，それらの関係を統計的な手法を用いて学習し，回帰モデルを導出する．入力画像には回帰モデルを適用し，姿勢推定を行なう．しかし，姿勢変化には周期性があり，これを考慮して回帰モデルを学習しなければ，推定精度が低下してしまう．

それに対して後者は，あらかじめ対象物体を様々な角度から撮影した大量の画像とのテンプレートマッチングを行ない，最も類似した画像が対応する姿勢を，推定結果として出力する．学習を必要としないため，姿勢変化の周期性の影響を受けにくい．しかし，多種多様な姿勢変化に対応するために，膨大な数のテンプレートを記憶しておく必要がある．

この問題に対し，Murase らは，3 次元姿勢変化による 2 次元画像上での見えの変化を，主成分分析によって導出した低次元空間における多様体で表現するパラメトリック固有空間法を提案した [4]．多様体構築に用いたテンプレートの系列をキュービックスプライン等で補間することで未知の姿勢にも対応できるため，記憶しておくべきテンプレートを減らすことができる．

しかし，主成分分析による多様体構築では画像全体の見えに着目するため，図 1 のように姿勢変化量が同じだとしても見えが似ている場合に，低次元空間上でほぼ同一の点に写像されてしまう場合がある．その様子を図 2 に示す．このような多様体では近い点に写像される姿勢を分離することが困難なため，姿勢推定精度が低下することが考えられる．これは，主成分分析は画像全体の見えを考慮した教師なし学習であり，姿勢の分離性について考慮していないことが原因である．

そこで我々は，姿勢の分離性に着目した特徴量による多様体構築手法を提案する．姿勢を教師信号として学習した Deep

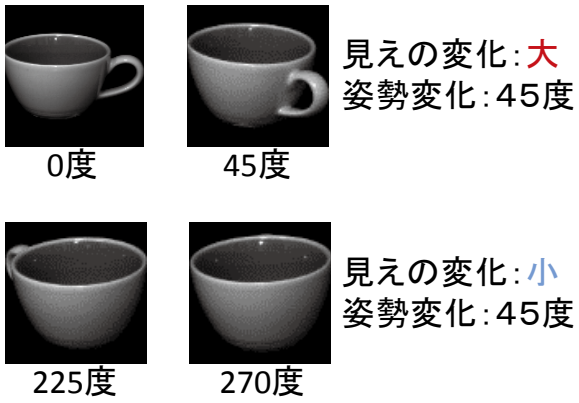

図 1　見えの変化と姿勢変化

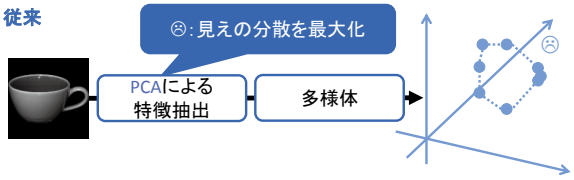

図 2　PCA による多様体構築

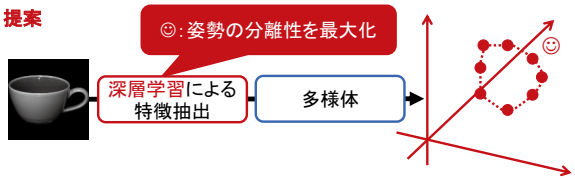

図 3　深層学習による多様体構築

Convolutional Neural Network（DCNN）[5] の中間層から，姿勢の分離性が高い特徴を抽出する．この特徴量を用いて多様体を構築する．その様子を図 3 に示す．DCNN は学習過程において識別に適した特徴量を自動的に獲得することができ，一般物体認識やシーン認識など様々なベンチマークで高い性能を示している [6]．そのため，図 1 に示すように，画像全体の見えの分散を最大化する教師なし学習である主成分分析では区別できない見えの変化が小さい姿勢の違いであっても，姿勢を教師信号として教師あり学習を行った DCNN を用いて抽出した姿勢の分離性の高い特徴量を用いることで区別することができる．

以下，2 節で多様体に基づく姿勢推定の概要について述べ，3 節で提案手法について述べる．4 節で提案手法の有効性について調査した実験と結果について述べ，考察を加える．最後に，5 節でまとめと今後の課題について述べる．

2. 多様体に基づく姿勢推定

多様体構築に基づく姿勢推定処理手順を図 4 に示す．多様体に基づく姿勢推定を行なうには，まず主成分分析等により物体の姿勢変化を特徴空間上で表現する多様体を構築する．姿勢推定には入力画像を多様体の存在する特徴空間へ写像し，多様体上の最も近い 1 点が示す姿勢を推定結果として出力する．

ここで，主成分分析は画像全体の見えを考慮した教師なし学習であるため，見えの変化が小さい姿勢の違いを区別すること

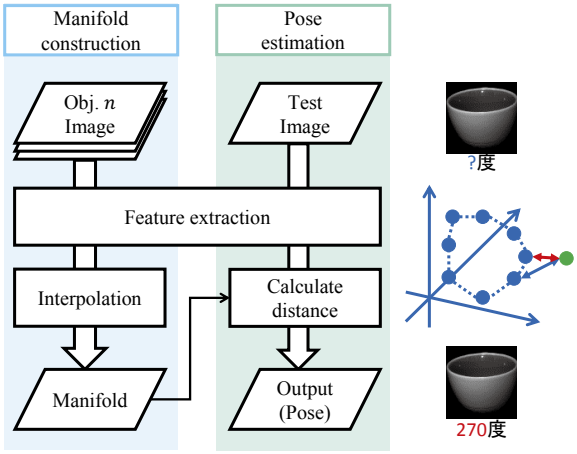

図 4　多様体構築に基づく姿勢推定処理手順

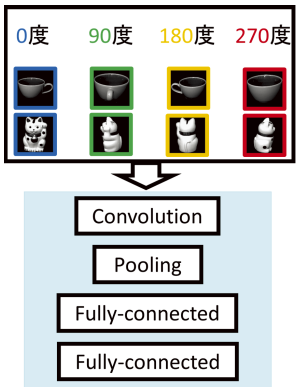

図 5　姿勢を教師信号とした DCNN の学習手順

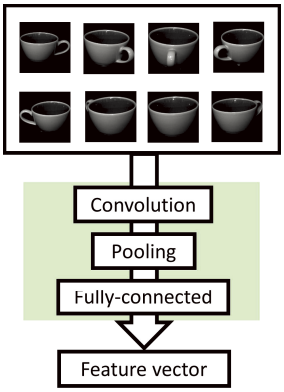

図 6　学習済み DCNN からの特徴抽出手順

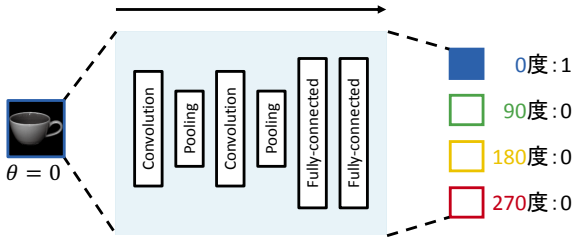

図 7　PoseNetC モデルの概要

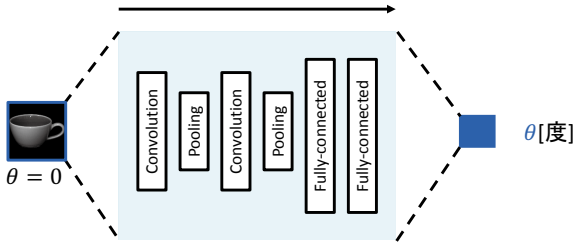

図 8　PoseNetR モデルの概要

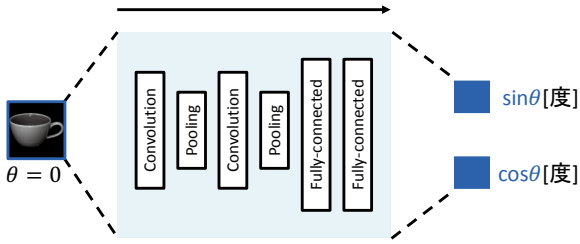

図 9　TriNetR モデルの概要

が困難である．そこで，多様体構築に姿勢を教師信号とした教師あり学習手法を用いることが考えられる．

その中で，我々は深層学習による特徴抽出に着目した．深層学習とは機械学習手法の 1 つであり，特徴抽出から識別までを一貫して学習できる手法である．これにより学習される特徴量は人手で設計された特徴量と比較して高い表現能力を持ち，他の識別器の学習においても有効であることが知られている [7]．また，物体の質感の認識のような，人手による特徴設計が難しい問題においても，深層学習により抽出された特徴量は高い性能を示す [8]．このことから，姿勢を教師信号として学習した深層学習モデルを用いることで，姿勢の分離性の高い特徴抽出ができると考えられる．そして，その特徴量を用いて構築した多様体ならば，主成分分析による多様体では対応できない見えの変化が小さい姿勢の違いを区別することができる．

深層学習を用いた多様体構築を行なうためには，まず深層学習モデルの 1 つである DCNN の学習を行う．学習サンプルには，推定対象物体を任意の回転軸に従って，一定の角度ごとに回転させた画像を用いる．ここで，図 5 に示すように，教師信号として各物体の姿勢を与える．これにより，姿勢の分離性に着目した学習が行われる．そして，学習した DCNN を用いて特徴抽出を行う．図 6 にその処理手順を示す．学習した DCNN に推定対象物体の学習サンプルを再び入力し，中間層の出力を特徴量として抽出する．このように抽出した特徴量に対し，パラメトリック固有空間法と同様に補間処理を行ない，多様体を構築する．

3.　姿勢を教師信号とした DCNN の学習

姿勢の分離性が高い特徴を抽出するために，姿勢を教師信号として DCNN を学習する．その際，教師信号の与え方の異なる以下の 3 つのモデルを提案する．

- PoseNetC：物体姿勢を教師信号とした分類モデル．
- PoseNetR：物体姿勢を教師信号とした回帰モデル．
- TriNetR：姿勢の 3 角関数値を教師信号とした回帰モデル．

3.1　PoseNetC

PoseNetC とは物体姿勢を教師信号とした分類モデルである．このモデルでは，姿勢を離散化し，姿勢推定問題をクラス分類問題として定式化する．

図 7 にモデルの概要を示す．全ての物体でその種類によらず，姿勢を教師信号として学習を行なう．つまり，姿勢 θ を離散化し，対応するクラスのみを 1，それ以外を 0 とした学習を行なう．そのため，Output layer の Unit 数は学習に用いた姿勢の種類だけ存在する．誤差関数には交差エントロピーを用いる．

このように学習を行なうことで，特徴空間上で同一クラスに属する姿勢がそれぞれ異なる 1 点に集中するような学習が行な

図 10　COIL-20 データセット中の推定対象物体

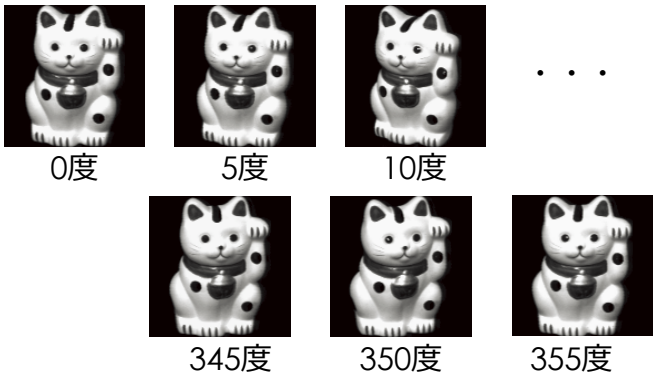

図 11　姿勢変化の例

われると考えられる．これにより，従来手法において見えが似ているため，特徴空間上でほぼ同一の点に存在していたものが分離されるような効果が得られると見込まれる．

3.2　PoseNetR

PoseNetR とは物体姿勢を教師信号とした回帰モデルである．

図 8 にモデルの概要を示す．PoseNetC と同様に，全ての物体で，その姿勢を教師信号として学習を行なう．しかしこちらは回帰モデルであり，姿勢 θ を連続値として扱う．そのため，Output layer の Unit 数は 1 つである．誤差関数には 2 乗誤差を用いる．

なお，このモデルは姿勢の周期性を考慮していないため，特に 0 度付近において，誤差が不当に大きくなる場合が考えられる．例えば，0 度の姿勢を 355 度と推定した場合，実際には 5 度の誤差であるが，355 度の誤差として学習してしまう．

3.3　TriNetR

TriNetR とは姿勢の 3 角関数値を教師信号とした回帰モデルである．

図 9 にモデルの概要を示す．PoseNetR と同様に回帰モデルであり，誤差関数には 2 乗誤差を用いる．このモデルでは姿勢の周期性を考慮するために，姿勢 θ を sin 関数と cos 関数で表現したものを教師信号とする．そのため，Output layer の Unit 数は 2 つである．

このモデルは姿勢の周期性を考慮した学習が行なわれるため，PoseNetR の問題点を解決できると考えらる．

表 1　DCNN のネットワークアーキテクチャ

Input	Units: 128 × 128
Convolution 1	Kernel: 5 × 5 Channel: 16 Maxpooling: 5 × 5
Convolution 2	Kernel: 5 × 5 Channel: 16 Maxpooling: 5 × 5
Fully-connect 3	Units: 512
Fully-connect 4	Units: 512
Fully-connect 5	Units: 512
Output	Units: 36 (PoseNetC) Units: 1 (PoseNetR) Units: 2 (TriNetR)

表 2　DCNN による姿勢推定評価値

	平均絶対誤差
PoseNetC	7.92 度
PoseNetR	28.32 度
TriNetR	9.29 度

4.　評 価 実 験

提案手法の有効性を確認するため，公開データセットを用いた物体の姿勢推定実験を行なった．なお，本報告では予備実験として鉛直軸周りの姿勢推定を行なった．

公開データセットには Columbia Object Image Library（COIL-20）[9] を用いた．推定対象物体は 20 種類で，それぞれについて鉛直軸に沿って 5 度刻みで姿勢を変化した画像があり，画像の総数は 1,440 枚である．データセット中の推定対象物体を図 10 に，姿勢変化の例を図 11 に示す．

4.1　DCNN による姿勢推定

多様体を用いない場合の姿勢推定精度を調査するため，深層学習モデルのみでの姿勢推定実験を行なった．PoseNetC, PoseNetR, TriNetR を比較した．各モデルのネットワークアーキテクチャを表 1 に示す．教師信号の与え方の違いから Output layer のみ Unit 数が異なっているが，その他はすべて同じ構造である．DCNN の Kernel や結合重みの初期値は乱数で決定した．活性化関数には Rectified Linear Units（ReLU）[10] を用いた．誤差関数には，分類モデルには交差エントロピーを，回帰モデルには 2 乗誤差を用い，誤差逆伝搬法にて Kernel や結合重みを更新した．なお，汎化能力を高めるために，Fully-connected layer の Unit を指定した割合だけランダムに選択し，その応答値を 0 にする Dropout [11] という処理を行なった．評価は学習サンプルを姿勢ごとに分割した 2 分割交差検定を行なった．つまり，以下に示すようにそれぞれ 10 度刻みの学習セットが存在する．

- セット 1 : 0 度，10 度，20 度，…，350 度
- セット 2 : 5 度，15 度，25 度，…，355 度

実験結果を表 2 に示す．PoseNetC が最も高精度であることがわかる．PoseNetR は姿勢の周期性を考慮せずに学習を行なっているため，他のモデルと比較して精度が低かった．

表 3　多様体による姿勢推定評価値

	平均絶対誤差
Pixel	1.16 度
PCA	1.39 度
ObjNetC	1.70 度
PoseNetC	1.09 度
PoseNetR	1.59 度
TriNetR	1.72 度
OverFeat [13]	1.89 度

TriNetR は姿勢の周期性を 3 角関数を用いることで表現してるため，PoseNetR と比較して精度が向上したが，PoseNetC と比較して精度が低かった．この原因として，

$$\sin\theta^2 + \cos\theta^2 = 1 \tag{1}$$

のような制約条件を学習時に与えていないため，姿勢の周期性を十分に表現できなかったことが考えられる．それに対し，PoseNetC は分類モデルであり姿勢の周期性が学習に与える影響が小さいため，高精度となった．しかし，分類モデルであるため，学習に用いなかった姿勢は推定できないという問題がある．今回の条件では，平均絶対誤差が 5 度を下回ることができない．

4.2　DCNN を用いた多様体構築による姿勢推定

深層学習を用いた多様体構築の有効性を調査するために，様々な特徴量で構築した多様体による比較を行なった．深層学習を用いない特徴量として，画素値をそのまま並べて特徴ベクトルとしたものと，そのベクトルの集合に対して主成分分析を行ない導出した固有ベクトルを用いた．深層学習を用いるものとして，PoseNetC，PoseNetR，TriNetR に加え，比較として物体種類を教師信号とした 2 つのモデルを用意し，それぞれから抽出した特徴量を用いた．物体種類を教師信号とした 2 つのモデルとして，COIL-20 データセットを用いて学習した ObjNetC モデルと，ImageNet 2012 training set [12] を用いて学習した一般物体分類モデルである OverFeat モデル [13] を用意した．

各深層学習モデルにおいて Convolution layer のすぐ後にある Fully-connected layer から特徴抽出を行なった．つまり，COIL-20 を用いて学習した各深層学習モデルは Fully-connect 3 から，OverFeat モデルは Fully-connect 8 から特徴抽出を行ない，それぞれ 512 次元，4,096 次元の特徴量を用いて多様体構築を行なった．主成分分析による多様体構築では，物体ごとに寄与率 80%以上となる次元数の固有ベクトルを特徴量とした．なお，固有ベクトルの次元数は平均 10 次元程度であった．画素値をそのまま並べたものを特徴量として用いた多様体構築では，16,384 次元特徴量を用いた．評価は 2 分割交差検定を行なった．なお，分割方法は DCNN による姿勢推定と同じである．

実験結果を表 3 に示す．PoseNetC から抽出した特徴量を用いた場合が最も高精度であることがわかる．ObjNetC や OverFeat などの物体種類を教師信号として学習された分類モデルは，姿勢を教師信号としていないことから姿勢の分離性を考慮した特徴量を学習することができず，精度が低かった．また，ObjNetC と OverFeat を比較して，一般物体で学習を行うよりも推定対象物体で学習したほうが精度が良いことがわかった．PoseNetR や TriNetR などの物体姿勢を教師信号として学習された回帰モデルは，姿勢を教師信号としているが，姿勢の周期性を十分に表現できなかったため，学習に悪影響を及ぼし，精度が低くなったと考えられる．それに対し，PoseNetC は物体姿勢を教師信号とした分類モデルであり，姿勢の周期性が学習に与える影響が小さく，姿勢の分離性が高い特徴量を学習できたため，高い精度を示したと考えられる．また，表 2 に示す DCNN による姿勢推定評価値と比較して，多様体を用いることで学習に用いなかった未知の姿勢も推定可能となるため，全ての特徴量において精度が向上した．

以上の結果より，深層学習を用いた多様体構築の有効性を確認した．

表 4　DCNN による姿勢推定評価値（Unit 数ごと）

	Units: 256	Units: 512	Units: 1024
PoseNetC	12.36 度	7.92 度	7.33 度
PoseNetR	32.42 度	28.32 度	27.00 度
TriNetR	12.76 度	9.29 度	9.75 度

表 5　多様体による姿勢推定評価値（Unit 数ごと）

	Units: 256	Units: 512	Units: 1024
ObjNetC	1.76 度	1.70 度	1.91 度
PoseNetC	1.17 度	1.09 度	1.16 度
PoseNetR	1.59 度	1.59 度	1.75 度
TriNetR	1.62 度	1.73 度	2.11 度

4.3　DCNN のネットワークアーキテクチャの違いが姿勢推定精度に与える影響

検討として，DCNN のネットワークアーキテクチャの違いが姿勢推定精度に与える影響の調査を行なった．ここでは層数を固定し，Fully-connected layer の Unit 数を変化させた．Fully-connected layer の Unit 数をそれぞれ 256，512，1024 で統一した 3 つの場合で，同様に姿勢推定実験を行った．

実験結果を表 4，5 に示す．DCNN の性能は Unit 数が増加するほど向上する傾向が見られるが，多様体による姿勢推定を行なう場合は Unit 数が 512 の場合が性能が高かった．DCNN の性能は中間数の Unit が増加するほど表現能力が増加するため向上したと考えられる．しかし，多様体構築に用いる場合にはそのような傾向は見られなかったため，最適な Unit 数を実験により求める必要がある．

5.　む　す　び

本報告では，深層学習を用いた多様体構築により物体姿勢を高精度に推定する手法について検討した．姿勢を教師信号として深層学習モデルを学習することで姿勢の分離性の高い特徴量を抽出し，その特徴量を用いて多様体を構築することで，見えの変化が小さい姿勢の違いも区別できる手法を提案した．評価実験より，多様体構築に主成分分析を用いる従来手法等と比較して，高精度に姿勢を推定できることを確認した．

今後の課題として，ネットワークの層数など，まだ検討でき

ていない学習パラメータについて調査を進めることや，複雑背景・照明変動に対する頑健性の調査等が挙げられる．

謝辞　本研究の一部は，科学研究費補助金による．

文　献

[1] N. Correll, K.E. Bekris, D. Berenson, O. Brock, A. Causo, K. Hauser, K. Okada, A. Rodriguez, J.M. Romano, and P.R. Wurman, “Lessons from the Amazon picking challenge,” ArXiv:1601.05484, Jan. 2016.

[2] J. Broekens, M. Heerink, and H. Rosendal, “Assistive social robots in elderly care: A review,” Gerontechnology, vol.8, no.2, pp.94–103, April 2009.

[3] M. Torki and A. Elgammal, “Regression from local features for viewpoint and pose estimation,” Proc. 2011 IEEE Computer Society Conf. on Computer Vision, pp.2603–2610, Nov. 2011.

[4] H. Murase and S.K. Nayar, “Visual learning and recognition of 3-D objects from appearance,” Int. J. Computer Vision, vol.14, no.1, pp.5–24, Jan. 1995.

[5] A. Krizhevsky, I. Sutskever, and G.E. Hinton, “ImageNet classification with deep convolutional neural networks,” Proc. 26th Annual Conf. on Neural Information Processing Systems, pp.1097–1105, June 2012.

[6] A.S. Razavian, H. Azizipour, J. Sullivan, and S. Carlsson, “CNN features off-the-shelf: An astounding baseline for recognition,” Proc. 2014 IEEE Conputer Society Conf. of Computer Vision and Pattern Recognition, pp.806–813, June 2014.

[7] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, “DeCAF: A deep convolutional activation feature for generic visual recognition,” arXiv preprint arXiv:1310.1531, Oct. 2013.

[8] M. Cimpoi, S. Maji, and A. Vedaldi, “Deep filter banks for texture recognition and segmentation,” Proc. 2015 IEEE Computer Society Conf. of Computer Vision and Pattern Recognition, pp.3828–3836, June 2015.

[9] S.A. Nene, S.K. Nayar and H. Murase, “Columbia Object Image Library (COIL-20),” Technical Report, Columbia Univ. Dept. of Computer Science, CUCS-005-96, Feb. 1996.

[10] V. Nair and G.E. Hinton, “Rectified linear units improve restricted Boltzmann machines,” Proc. 27th Int. Conf. on Machine Learning, pp.807–814, June 2010.

[11] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R.R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” arXiv preprint arXiv:1207.0580, Oct. 2012.

[12] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” Proc. 2009 IEEE Computer Society Conf. of Computer Vision and Pattern Recognition, pp.248–255, June 2009.

[13] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus and Y. LeCun, “Overfeat: Integrated recognition, localization and detection using convolutional networks,” arXiv preprint arXiv:1312.6229, Dec. 2013.