Vol.2017-CVIM-207 No.12
2017/5/10

深層学習による局所画像特徴量を用いた 3 次元形状比較

小林万希子 [1,a)]　古屋貴彦 [1,b)]　大渕竜太郎 [1,c)]

概要：3 次元形状モデル (3D モデル)の類似比較では，3D モデルの形状表現，幾何変形 (平行移動，拡大縮小，回転)，姿勢変化や大域的変形に対する不変性が要求される．これら不変性の獲得をねらったアプローチとして，見かけの局所特徴量集合を用いる手法がある．Furuya ら [2]は，3D モデルの多視点見かけ画像から局所画像特徴量を多数抽出し，それら局所特徴を 3D モデル当たり 1 個の特徴量に統合することで，3D モデルを効率良く比較した．しかし Furuya らの手法は人手で設計された局所画像特徴量と統合法を用いるため，形状比較に最適な統合特徴量が得られるとは限らない．本研究では，見かけに基づく 3D 形状比較の精度改善をねらい，深層畳み込みニューラルネットワークを用いて形状比較に適した局所画像特徴量とそれらの統合法をデータ駆動で学習する．提案手法を 3D モデル検索のシナリオで評価した結果，提案手法は state-of-the-art な手法には及ばないものの，Furuya らの手法より高い精度を示すことが分かった．

キーワード：深層学習，畳み込みニューラルネットワーク，局所視覚特徴量，Bag-of-Features，3 次元形状検索

1. はじめに

近年，安価な 3D レンジスキャナや 3D プリンタ等の普及や共有ウェブサイトの登場等により，3D モデルの数が爆発的に増加している．3D モデルの用途も拡大しており，例えば，3D 機械 CAD を用いた工業製品の設計，3D 分子構造の解析による製薬，MRI で撮像された 3D 臓器画像からの病気診断，など幅広い．3D モデルの効率的な再利用や 3D 形状の効果的な解析のためには，3D モデルをその形状の類似性に基づいて比較，検索，分類等を行う技術が必要である．

3D 形状の比較では，図 1 に示すように，3D モデルの (1) 形状表現（点群，ボクセル，ポリゴン等），(2) 幾何変換（回転，平行移動，一様スケーリング），(3) 大域的変形や姿勢変化，に対する不変性が要求される．さらに，多数の 3D モデルを含む大規模データベースにおいては，コンパクトな 3D モデル特徴量と高速な特徴比較が重要である．

これまでに，数多くの 3D モデル検索手法が提案されてきた [1]. それらの中でも，見かけの局所特徴集合とその統合を用いるアプローチ (例えば，[2, 3])は，上記の不変性と高い計算効率を持つ．Furuya らの BF-DSIFT 法 [2]は，3D モデルの多視点見かけ画像から，画像面内の回転に頑強なマルチスケール局所画像特徴量を多数抽出し，これら局所特徴量を Bag-of-Features (BF)法 [4]で 3D モデル当たり 1 個の特徴量に統合する．見かけによる比較は 3D モデルの形状表現に依存しない．また，スケール変化と回転に頑強な局所特徴は 3D モデルの幾何変形と姿勢変化に対する頑強性を向上させる．さらに，BF 法による統合は，3D モデルの比較コストを大幅に低減する．

(1) 形状表現　(2) 幾何変換　(3) 姿勢変化

図 1　3D 形状比較に要求される不変性.

BF-DSIFT 法は 3D モデル検索の国際コンテストで検索精度上位の成績を収めた [5, 6]が，その精度は実用には未だ不十分である．精度が低い主な理由は，BF-DSIFT 法が人手で設計された (手作りの)局所画像特徴量と統合法を用いることである．BF-DSIFT 法が用いる SIFT [7]は本来，自然画像の部分マッチングのために設計された局所画像特徴量であり，3D 形状の特徴記述に適するとは限らない．また，BF 法で得た統合特徴量が 3D 形状の比較に向かない可能性もある．

本研究の目的は，形状表現，幾何変換，姿勢変化に対する不変性と高い計算効率を備え，かつ，先行研究よりも高精度な 3D 形状比較である．我々は深層学習を導入し，3D 形状比較に最適な局所画像特徴量の抽出とそれらの統合をねらう．我々が提案する *Local Visual feature Aggregation Network* (*LVAN*)は，3D モデルの多視点見かけ画像群から切り出した多数の局所画像からの特徴抽出，および，局所特徴量群の統合を 1 つの Deep Convolutional Neural Network (DCNN)で実現する．3D モデルの幾何変形に対する頑強性を高めるために，局所画像の大きさを多様化し，かつ，局所画像に対して画像面内の回転正規化を施す．

LVAN を効果的に学習するには，非常に多数のラベル付き 3D モデルが必要である．しかし，2 次元自然画像と比べて，ラベル付き 3D モデルの収集は困難である．また，LVAN と 2 次元自然画像向けの DCNN (例えば，[8][9])とではネッ

1 山梨大学
University of Yamanashi
a) g17tk007@yamanashi.ac.jp,
b) takahikof@yamanashi.ac.jp
c) ohbuchi@yamanashi.ac.jp

トワーク構造が異なることから，自然画像で事前に訓練された既存 DCNN を転移することも現実的でない．そこで本研究では，データ拡張と 2 段階学習を行うことで，比較的少数 (~10,000 個)のラベル付き 3D モデル群から効果的に LVAN を学習する．

3D モデル検索の標準ベンチマークを用いた評価実験の結果，LVAN は state-of-the-art な検索手法には及ばないものの，BF-DSIFT より有意に高い検索精度を示すことが分かった．また，LVAN の 2 段階学習が精度改善に効果的であることを確認した．

2. 関連研究

3D モデルからの形状特徴抽出の手法は様々ある．近年では，DCNN を利用した 3D 形状比較のための研究が盛んである (例えば，[10, 11, 12, 13, 14, 15])．DCNN から抽出された 3D 形状特徴量は一般的に，従来の手作り特徴に比べて高精度である．本章では，BF-DSIFT 法に加え，DCNN を用いて 3D 幾何特徴量を抽出する DLAN 法 [10]，DCNN を利用した局所画像特徴量 LIFT [16]について述べる．

2.1 BF-DSIFT

BF-DSIFT [2]は，形状表現，幾何変換，姿勢変化に対する不変性の獲得と，形状比較の計算コストの低減をねらって提案された．図 3 に BF-DSIFT 法による特徴抽出の流れを示す．まず，3D モデルを複数の視点 (例えば 42 視点)からレンダリングして得た深さ画像から，局所画像特徴量を密に多数 (例えば画像当たり 300 個)抽出する．局所画像特徴量には，画像の照明変化，スケール変化，回転変化に頑強な Scale Invariant Feature Transform (SIFT)特徴量 [7]を用いる．注意点として，BF-DSIFT は 3D 形状の多様な特徴を捉えるために，SIFT を抽出すべき局所画像領域を，顕著点検出法ではなく，密かつランダムなサンプリングにより決定する．画像面内における回転に対する頑強性を得るため，密ランダムサンプリングされた局所領域は，SIFT の方向推定法を用いて回転正規化される．推定される向きは，局所領域内の輝度勾配が最大となる向きと概ね一致する．

上記の処理の結果，1 個の 3D モデル当たり約 13,000 個の SIFT 特徴量が抽出される．次いで，BF 法を用いて，これら SIFT 特徴量群を 3D モデル当たり 1 個の特徴量に統合する．3D モデル 1 対の類似度は，それら 3D モデルの統合特徴量を比較することで計算される．

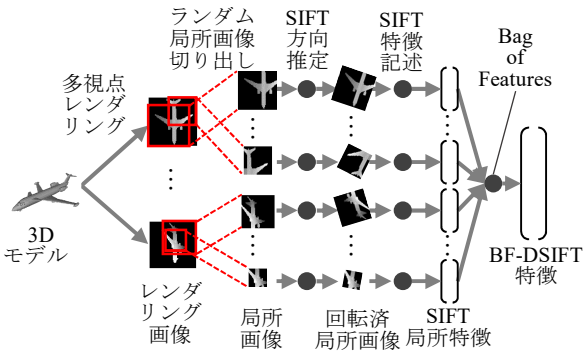


図 3 BF-DSIFT 法 [2]による特徴抽出．

2.2 LIFT

Kwang ら [16]は，手作りの SIFT 特徴量のマッチング精度を改善させるため，DCNN を用いた局所画像特徴量 Learned Invariant Feature Transform (LIFT)を提案した．特徴抽出の流れは SIFT [7]と類似する．即ち LIFT は，局所特徴量を抽出すべき局所領域 (顕著点)の検出，局所領域の方向推定，局所領域の特徴記述の 3 処理から成る．SIFT の各処理が人手で設計されているのに対し，LIFT は DCNN を用いて，各処理をデータ駆動で学習する．LIFT の DCNN は，顕著点検出に 1 層，方向推定に 5 層，特徴記述に 3 層の畳み込み層を持つ．LIFT の学習には自然画像が用いられる．学習用データは，様々な

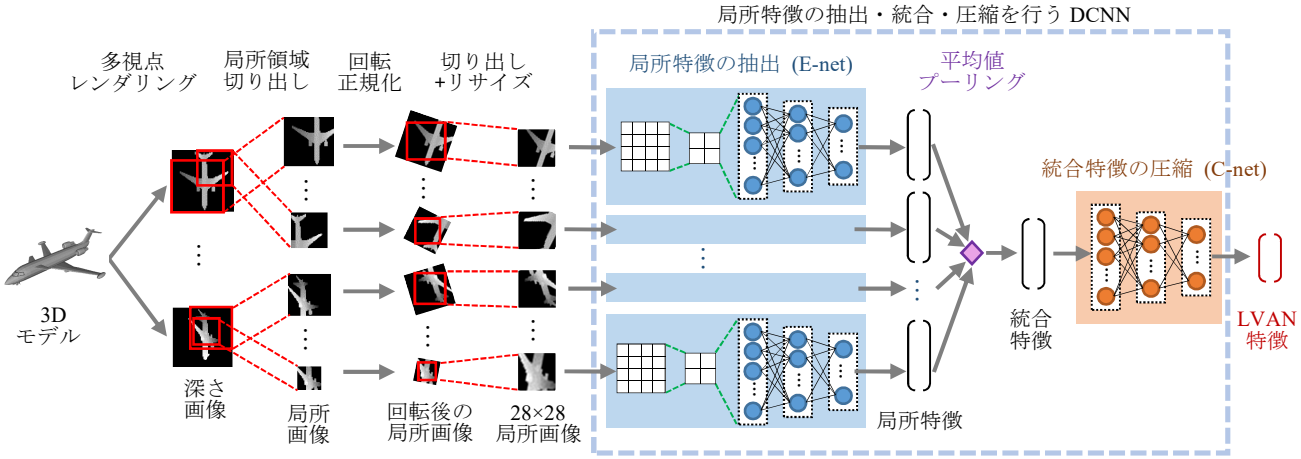


図 2　LVAN 法は，3D モデルの見かけ画像から多様な領域を切り出し，向きを正規化した局所画像を DCNN で特徴記述する．局所特徴量群は平均値プーリングにより統合され，後続の DCNN でさらにコンパクトかつ顕著な特徴量へ圧縮される．

角度から撮影された画像 [17]や，様々な時間帯に撮影された画像[18]である．特徴記述 DCNN は，埋め込み特徴空間上での相対関係に基づいて訓練される．即ち，埋め込み特徴空間において，同じ顕著点を撮影した画像から抽出される特徴量は互いに近く，異なる顕著点を撮影した画像から抽出される特徴量は互いに遠くなるように訓練される．

LIFT は局所画像特徴量を最適化する一方で，それらの統合法は学習しない．本論文における実験では，3D モデルの多視点見かけ画像から LIFT 特徴量を抽出し，それらを BF 法で統合した BF-LIFT 法の検索精度を，提案手法と比較する．図 3 に BF-LIFT 法による特徴抽出の流れを示す．

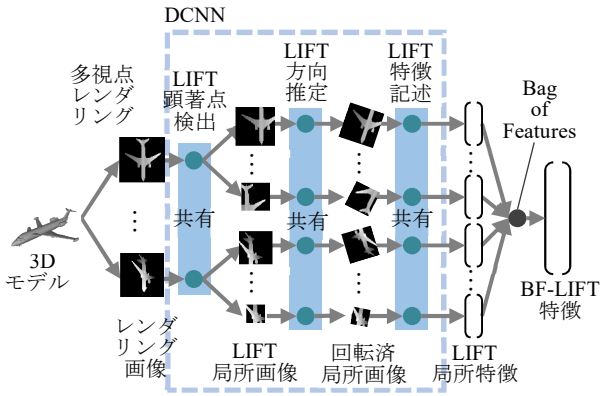


図 4 BF-LIFT 法による特徴抽出．

2.3 DLAN

Furuya ら [10]は，3D の DCNN を利用して 3D モデルから 3D 幾何特徴量を抽出する Deep Local feature Aggregation Network (DLAN)を提案した．図 5 に DLAN 法による特徴抽出の流れを示す，DLAN は 3D 有向点群モデルから多様な位置，大きさの 3D 局所領域を複数 (3D モデル当たり 100 個程度)サンプリングし，各局所領域に含まれる点群の分布統計量に基づいて局所領域を回転正規化する．局所領域は 3D 空間情報付きの低レベル特徴量で記述され，これを 3D の DCNN へ入力し，特徴精製する．精製された局所特徴量はネットワークの中間層で 3D モデル当たり 1 個の特徴量に統合される．DLAN は 3D モデル検索における state-of-the-art な手法群の中の 1 つである．

我々は，DLAN 法のアイデア (即ち，DCNN を用いて，3D モデルから多数の局所領域を切り出し，局所特徴量の精製，統合を行う)を 2D の見かけに基づく 3D 形状比較に適用する．

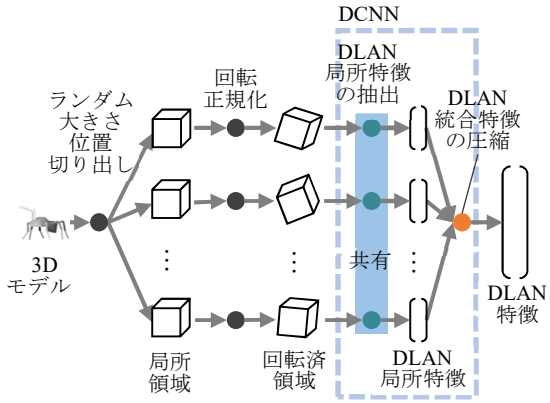


図 5 DLAN 法 [10]による特徴抽出．

3. 提案手法

提案手法である LVAN は，3D モデルからコンパクトかつ高精度な見かけ特徴量を抽出する．図 2 に，LVAN による特徴抽出処理の流れを示す．LVAN は，3D モデルからの局所画像特徴量群の抽出，それらの統合，および，統合特徴の圧縮を行う DCNN である．統合特徴を圧縮して得た LVAN 特徴量を比較することで，3D モデル間の類似度を計算する．3.1 節では LVAN の構造と，LVAN 特徴量を用いた 3D モデル検索について説明し，3.2 節では LVAN の効果的な学習手法を説明する．

3.1 LVAN 法による特徴抽出

LVAN の特徴抽出は，(1) 局所画像の生成，(2) 局所特徴抽出，(3) 局所特徴量の統合，(4) 統合特徴の圧縮，の 4 つの処理から成る．(2)～(4)の処理は DCNN で実現される．表 1 に DCNN の構成を示す．

(1) 局所画像の生成

3D モデルの多視点 (42 視点)の見かけ画像を得る．このためにまず，3D モデルの位置と大きさを正規化する．具体的には，3D モデルの頂点の重心が 3D 空間の原点と重なるように 3D モデルを平行移動し，次いで原点から最も遠い頂点の長さが 1 になるように 3D モデルを一様スケーリングする．位置と大きさを正規化した 3D モデルを 80 面体で囲み，80 面体の頂点 (42 個)の各々に視点を設置する．各視点の視線は原点 (即ち 3D モデルの重心)に向ける．各視点から 3D モデルを z-buffer レンダリングすることで，42 枚の深さ画像を得る．レンダリング解像度は 256×256 pixel とする．図 6 に，飛行機の 3D モデルを多視点レンダリングして得た深さ画像の例を示す．

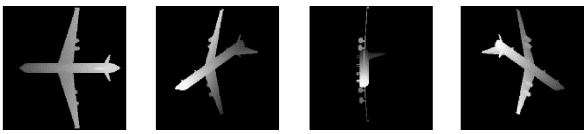


図 6 3D モデルの多視点深さ画像の例 (airplane カテゴリ)．

深さ画像の各々から局所画像を切り出す．深さ画像から領域をランダムな位置，大きさで選択し，正方形を切り出す．画像面内の回転に対する頑強性を獲得するために，局所画像の回転正規化を行う．本研究では，輝度値を用いて局所画像の重心を計算し，その重心と局所画像の中心とを結ぶベクトルの向きが全ての局所画像間で同じになるように，各局所画像を回転させる．回転により生じる物体の「切れ目」が含まれないよう，回転した局所画像に内接する正方形を再度切り出す (図 2)．作成された局所画像に 3D モデルが画像面積の 10%以上映っていないものは除外し，局所画像とする．切り出した局所画像をさらに 28×28 pixel にリサイズし，後続の DCNN へ入力する．

事前実験により，深さ画像 1 枚当たり局所画像枚数は 5 に決定した．レンダリング視点数が 42 であるため，1 つの 3D モデルから 42×5=210 枚の局所画像が切り出される．

(2) 局所特徴抽出

局所画像 1 枚に対して 1 つの局所特徴を抽出する．局所特徴抽出には DCNN (E-Net と表記)を用いる．表 1 に E-Net の構造を示す．E-Net は 2 つの畳み込み層と 1 つの全結合層から成る．DCNN 学習の収束を早めるため，畳み込み層には batch normalization [19]を適用し，各層の活性化関数には ReLU [8]を用いる．全結合層のユニット活性 (512 次元)を局所画像特徴量として用いる．

表 1　LVAN のネットワーク構成.

層		フィルタサイズ	ユニット数	pooling	活性化関数
E-Net	畳み込み	3×3	32	Max	ReLU
	畳み込み	3×3	64	Max	ReLU
	全結合	－	512	－	ReLU
統合		－	－	Average	－
C-Net	全結合	－	256	－	ReLU
	全結合	－	128	－	ReLU

(3) 局所特徴量の統合

E-Net から出力された局所特徴量 (512 次元)の集合を，平均値プーリングにより 3D モデル当たり 1 個の特徴量 (512 次元)に統合する. 別の統合法としては最大値プーリングがあるが，平均値プーリングは全ての局所特徴量が統合特徴量に等しく寄与する．局所画像のランダムサンプリングと，局所特徴量の平均値プーリングを組み合わせて用いることで, 3D モデルのあらゆる見かけ特徴を統合特徴へ反映させる．また，平均値プーリングは (最大値プーリングも同様に)，LVAN 全体 (E-Net と C-Net)を誤差逆伝播法で学習可能な利点がある．

(4) 統合特徴量の圧縮

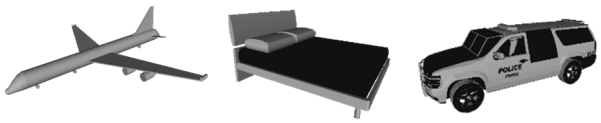
統合特徴量 (512 次元)をさらに圧縮し，コンパクトかつ高精度な特徴量を得る．統合特徴量の圧縮には全結合型ニューラルネットワーク (C-Net と表記)を用いる. 表 1 に C-Net の構造を示す．C-Net は 2 層の全結合層から成る．最後の全結合層から取り出した 128 次元の特徴ベクトルが LVAN 特徴である．

3D モデル検索を行う際は，検索要求として与えられた 3D モデルの LVAN 特徴量と, 検索対象 3D モデル群の LVAN 特徴量群とをユークリッド距離で比較する．検索要求との距離が小さい順に 3D モデルを並べ替え，検索結果としてユーザへ提示する．

3.2 LVAN の効果的な学習

本研究では 3D モデル検索のベンチマークである ModelNet40 [11]の training set を用いて LVAN を学習する．Training set は，飛行機，車，木など 40 個の物体カテゴリに分類された 9,843 個のカテゴリラベル付き 3D モデルを含む. 図 7 に, ModelNet40 に含まれる 3D モデルの例を示す.

しかしながら，一般的に，高々10,000 個のデータは DCNN を学習するには少数である．我々は，少数のラベル付き 3D モデルを用いて LVAN を効果的に訓練するため，2 段階の学習法を提案する．



(1) airplane　(2) bed　(3) car

図 7 ModelNet40 training set [11]に含まれる 3D モデル例.

(1) 1 段階目：E-Net の教師あり事前学習

局所特徴量を抽出するための E-Net を，多数の「ラベル付き局所画像」を用いて学習する. ラベル付き局所画像は，ModelNet40 training set に含まれるラベル付き 3D モデルから，3.1 節(1)の方法を用いて生成される．局所画像のラベルは，切り出し元の 3D モデルのラベルに等しい．本操作は，E-Net を訓練するための教示データの拡張を行うことに相当する．ラベル付き局所画像群は 3D モデルの多様な部位の多様なスケールの見かけ画像で構成されており，これを用いて E-Net を訓練することで，形状の幾何変形や多様性に対する頑強性を獲得させる．本研究では 9,843 個のラベル付き 3D モデルから，合計で 2,067,030 枚のラベル付き局所画像を生成する．

E-Net を事前学習するために，E-Net の全結合の次に softmax 関数の識別層を連結する．識別層のユニット数は 40 であり，物体カテゴリ数と等しい．目的関数は交差エントロピー誤差とし，最適化アルゴリズムには Adam [20]法 (初期学習係数は 0.001)を用いる．ミニバッチサイズは 17 とし，学習は 100 epoch 反復させる．

(2) **2 段階目：ネットワーク全体の教師あり学習**

1 段階目の学習が終了後, E-Net と C-Net を含むネットワーク全体の学習を行う．学習サンプルには，ModelNet40 training set に含まれるラベル付き 3D モデル (9,843 個)を用いる．1 段階目に用いた E-Net の識別層を削除し，C-Net の最後の全結合層の次に識別層を連結する．識別層のユニット数は 40 である．E-Net のパラメタは 1 段階目の学習結果を継承し，C-Net のパラメタはランダムに初期化される．

1 段階目の学習と同様，交差エントロピー誤差を Adam 法で最小化する．初期学習係数は 0.03，ミニバッチサイズは 17，学習反復回数は 1,000 epoch 程度である．

4. 評価実験

提案手法の有効性を, 3D モデルの形状類似検索のシナリオで評価する. 実験では, 2 段階学習の効果の有無, ハイパーパラメタ (局所画像枚数等)と精度の関係, 学習回数とカテゴリ正解率の推移を調査する. また, 既存の 3D モデル検索手法群との精度比較を行う.

4.1 実験条件

評価用データセット：ModelNet40 [11]の test set を精度評価に用いる．ModeNet40 test set は LVAN の学習に使用した training set と同じ 40 個の物体カテゴリに分類された 2,468 個の 3D モデル群から成る. 検索精度の評価実験では, test set の 2,468 個の 3D モデルの中から 1 個の検索要求 3D モデルを選び，残りの 2,467 個の 3D モデル群を検索対象とする．2,468 個の 3D モデルの各々を検索要求とした場合の検索精度を算出し，全ての検索要求の平均値を，ModelNet40 test set における検索精度とする．

検索精度の評価尺度には Mean Average Precision (MAP) を用いる．MAP [%]は，3D モデルの検索結果 (順位リスト) を頭から走査し，適合モデルが得られた時点における適合率の平均を Average Precision として得る．MAP は，全 3D モデルの Average Precision の平均値である．

比較対象：検索精度の主な比較対象として, BF-DSIFT 法 (2.1 節)と BF-LIFT 法 (2.2 節)を用いる. BF 統合に用いるコードブックの語彙数は BF-DSIFT では 10, 000, BF-LIFT では 500 とする. BF-DSIFT には Furuya らの実装を用い, BF-LIFT の LIFT 特徴抽出には, Kwang らによる学習済みモデル [21]を用いる. この他, DCNN で獲得した 3D モデル特徴量を検索に用いる state-of-the-art な手法群 [10, 11, 12, 13, 14, 15]との精度比較も行う．

実験環境:実験には 2 つの PC を用いる．1 つ目の PC のハードウェア構成は，Intel Xeon CPU E5-2650 v2 @2.60GHz ×2, メインメモリ 256GB, NVIDIA GeForce GTX TITAN X, GPU メモリ 12GB である. 2 つ目は, Intel Xeon CPU E5-2680 v2 @2.80GHz×2，メインメモリ 128GB，NVIDIA GeForce GTX 1080，GPU メモリ 8GB である．LVAN は TensorFlow ライブラリ [22]を用いて実装した．学習に要した時間は 1 段階目の額十には 8 時間，2 段階目の学習には 3 日(2 つの PC で概ね同じ)である．

4.2 実験結果

4.2.1 視点当たりの局所特徴数と検索精度

表 2 に，LVAN 法における 1 視点当たりの局所画像枚数と検索精度の関係を示す．学習に要する時間と GPU メモリの上限を考慮し，1 視点当たり 5, 10, 20 枚の 3 通りについて実験した．本実験は LVAN の初期検討時に行ったものであり，中間層での batch normalization は用いない．

表 2 より, 視点当たりの局所画像枚数を増やすほど検索精度が改善することが分かる．より多くの部位を DCNN に「見せる」ことで，3D モデルの形状を正確に記述できる．視点当たり 20 枚より多くの局所画像を用いる実験は，学習時間とメモリ消費の面から困難であった．以降の実験では，LVAN の学習と評価を現実的な時間内で行うために，視点当たりの局所画像枚数を 5 に固定し実験する．

表 2　視点当たりの局所画像枚数と検索精度.

視点当たりの局所画像枚数	MAP[%]
5	48.0
10	48.4
20	52.0

4.2.2 2 段階学習の効果

2 段階学習の効果を確かめるために，(1) 2 段階学習を行う場合(「事前学習あり」と表記)と，(2) 1 段階目の学習を行わず，ネットワーク全体の学習のみを行う場合(「事前学習なし」と表記)，の 2 通りの精度を比較する．事前学習なしの場合は，C-Net の最後の全結合層の次に識別層を連結し，VLAN 全体のパラメタを乱数で初期化してから訓練する．

図 8 に, 学習反復回数と検索精度の関係を示す. 図 8 では, 多少の MAP 値の増減はあるものの，「事前学習あり」は「事前学習なし」よりも一貫して高い検索精度を示すことが見て取れる．多数のラベル付き局所画像を用いて，E-Net が 3D モデルの部分を見分けるように訓練することで，高精度な局所画像特徴量が得られた．

図 9 に，学習反復回数とカテゴリ正解率の関係を示す. 図 9 のグラフは，事前学習ありとなし，training set と test set，の全ての組み合わせ (4 通り)の推移を示す．正解率の推移は, 事前学習の有無による違いがほとんど見られない. 興味深いことに，事前学習は検索精度の改善に効果がある一方で，識別精度 (カテゴリ正解率)には影響をほぼ与えない．2 段階学習法が softmax 関数によるカテゴリ分類には影響を与えない範囲で統合特徴量の空間を変形した結果，より良い特徴間距離が得られるようになったと推察する．

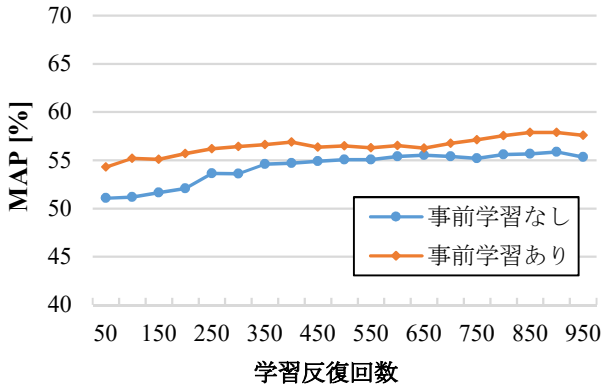


図 8 LVAN の学習反復回数と検索精度.

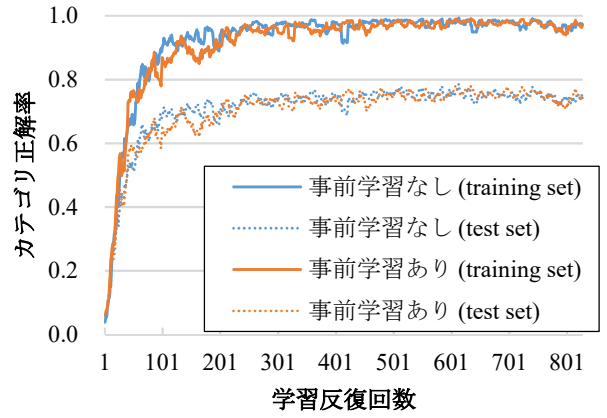


図 9 LVAN の学習反復回数とカテゴリ正解率.

4.2.3 既存手法との検索精度比較

表 3 に，提案手法を含む 9 つの 3D モデル検索手法の精度を示す．提案手法 LVAN の精度 (MAP=59.2 [%])は，BF-DSIFT の精度 (MAP=47.6 [%]) と BF-LIFT の精度 (MAP=31.8 [%])より高い．LVAN が BF-DSIFT と BF-LIFT よりも 3D 形状比較に適した統合特徴量を生成した．

BF-LIFT は局所特徴量の抽出に DCNN を用いるにも関わらず低精度である．LIFT の学習に用いられた自然画像は，3D モデルの見かけ画像とのデータの領域 (ドメイン)が異なる．そのため，自然画像で訓練された LIFT が 3D モデルの見かけからの顕著点検出，方向推定，特徴記述に失敗した可能性がある．

LVAN の検索精度は DCNN ベースの既存手法 (DLAN, MVCNN 等)には劣る結果となった. LVAN の DCNN は自然画像向けの DCNN (例えば[9])と比べて浅く，学習モデルとしての表現力が不十分だった可能性がある．LVAN の構造をさらに深層化したり，視点当たりの局所画像を増やしたりする (4.2.1 節)ことで精度改善が見込める．

表 3 ModelNet40 test set における検索精度.

手法	MAP[%]
LVAN（提案手法）	59.2
BF-LIFT	31.8
BF-DSIFT [2]	47.6
DLAN [10]	85.0
3D ShapeNets [11]	49.2
MVCNN [12]	79.5
Geometry Image [13]	51.3
GIFT [14]	82.0
DeepPano [15]	76.8

5. まとめと今後の課題

見かけの局所特徴量とその統合を用いた 3D 形状比較法は，3D モデルの形状表現，幾何変換，姿勢変化に対する頑強性が高く，かつ，比較の計算効率も高い．本稿では，この比較法の精度をさらに改善させるため，深層学習を用いて高精度な局所画像特徴量の抽出と統合を試みた．提案した Local Visual feature Aggregation Network (LVAN)は，3D モデルの多視点の見かけから切り出した部位からの特徴抽出，および，局所特徴量群の統合・圧縮を 1 つの DCNN で実現する．少数のラベル付き 3D モデルの形状を効果的に学習するため，LVAN の 2 段階学習法を提案した．

3D モデル検索のシナリオで提案手法を評価した結果, 提案手法は，state-of-the-art な手法には及ばないものの， BF-DSIFT を上回る検索精度を示した．この結果は，DCNN が手作りの局所画像特徴量と統合法よりも 3D 形状比較に適した統合特徴量を生成したことを示す．また，LVAN の 2 段階学習の有効性も認められた．

今後の課題は LVAN のさらなる高精度化である．現状の LVAN の検索精度は state-of-the-art な手法群と比べると低いが，改善の余地もある．例えば，LVAN に入力する局所画像の方向推定には単純な画素値統計を用いており，方向推定に失敗している可能性がある. LIFT 法のような方向推定 DCNN の導入が考えられる．また，LVAN のネットワーク構造については未探索であり，さらなる深層化により精度が改善する可能性がある. さらには, 見かけ (例えば LVAN) と 3D 幾何 (例えば DLAN)を組み合わせた形状比較などを検討する．

参考文献

[1] ElNaghy, H., Hamad, S., and Khalifa, M.E.. Taxonomy for 3D content-based object retrieval methods, International Journal of Research and Reviews in Applied Sciences, 14(2), pp.412–446, 2013.

[2] Furuya, T. and Ohbuchi, R.. Dense sampling and fast encoding for 3D model retrieval using bag-of-Visual features. ACM International Conference on Image and Video Retrieval 2009,

Article No. 26.
[3] Ohbuchi, R., Osada, K., Furuya, T., and Banno, T.. Salient local visual features for shape-based 3D model retrieval, IEEE Shape Modeling International 2008, pp. 93-102.
[4] Csurka, G., Bray, C., Dance, C. and Fan, L.: Visual categorization with bags of keypoints, Workshop on Statistical Learning in Computer Vision,European Conference on Computer Vision, pp.1–22 (2004).
[5] B. Li, et al.. SHREC'12 Track: Generic 3D Shape Retrieval, Eurographics Workshop on 3D Object Retrieval 2012, pp.119-126, 2012.
[6] Z. Lian, et al.. SHREC’10 Track: Non-rigid 3D Shape Retrieval, Eurographics Workshop on 3D Object Retrieval 2010, pp.101-108, 2010.
[7] Lowe, D. G.. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, 60 (2), pp. 91-110.
[8] Krizhevsky, A., Sutskever, I., and Hinton, G.E.. ImageNet Classification with Deep Convolutional Neural Networks, Advances in Neural Information Processing Systems 25 (NIPS 2012).
[9] Simonyan, K. and Zisserman, A.. Very Deep Convolutional Networks for Large-Scale Visual Recognition, arXiv technical report, 2014.
[10] Furuya, T. and Ohbuchi, R.. Deep Aggregation of Local 3D Geometric Features for 3D Model Retrieval. British Machin e Vision Conference 2016.
[11] Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., and Xiao J.. 3D ShapeNets: A Deep Representation for Volumetric Shape Modeling, Computer Vision and Pattern Recognition 2015.
[12] Su, H. and Maji, S. et al.. Multi-view Convolutional Neural Networks for 3D Shape Recognition. International Conference on Computer Vision 2015.
[13] Sinha, A. and Bai, J. et al.. Deep Learning 3D Shape Surfaces Using Geometry Images. European Conference on Computer Vision 2016, vol 9910, pp. 223-240.
[14] Bai, S. and Bai, X. et al.. GIFT: A Real-time and Scalabe 3D Shape Search Engine. Computer Vision and Pattern Recognition Conference 2016, pp. 5023-5032.
[15] Shi, B. and Bai, S. et al.. DeepPano: Deep Panoramic Representation for 3-D Shape Recognition. Signal Procession Letters 2015, 22(12), pp.2339-2343.
[16] Yi, K. M. and Trulls, E., Lepetit, V. et al.. LIFT: Learned Invariant Feature Transform. European Conference on Computer Vision 2016.
[17] Wilson, K. and Snavely, N.. Robust Global Translations with 1DSfM. European Conference on Computer Vision 2014.
[18] Verdie, Y. and Yi, K. M. et al.. TILDE: A Temporally Invariant Learned DEtector. Computer Vision and Pattern Recognition Conference 2015.
[19] Ioffe, S. and Szegedy, C.. Batch Normalization: Accelerating deep network training by reducing internal covariate shift. International Conference on Machine Learning 2015.
[20] Kingma D.P. and Ba, J.. Adam: A Method for Stochastic Optimization. International Conference on Learning Representations 2015.
[21] “cvlab-eplf/LIFT”. https://github.com/cvlab-epfl/LIFT, (参照 2017-04-10).
[22] Abadi, M. et al.. TensorFlow: A system for large-scale machine learning, USENIX conference on Operating Systems Design and Implementation 2016, pp. 265-283.