Vol.2019-CG-175 No.3
2019/9/28

PointNet#:色情報と時系列情報を付与した3次元点群の分類

PointNet#: Classification of time-series colored point sets

石塚唯矢 [1,2]　河合継 [1*]　眞嶋啓介 [1,3]

[1] クリスタルメソッド株式会社
[1] Crystal method Co. Ltd.
[2] 早稲田大学人間科学部
[2] Waseda University School of Human Sciences
[3] 慶応義塾大学環境情報学部
[3] Keio University Environment and Information Studies

Abstract: 近年 VR や AR,MR など 3D を扱う技術が発展を遂げ、「デジタルツイン」と呼ばれる現実空間の情報をセンサーで取得し仮想空間上に 3D で再現してシミュレーションや実験を行う取り組みが行われている。これに伴い、深層学習を用いた 3D データに対する分類や異常検知といった技術の需要が高まっている。3D 点群情報扱う深層学習モデルである Pointnet#[3] では、色情報が付与された点群の分類が行えるようになった一方で、時間情報が付与された点群を学習させる研究は行われていない。そこで本論文では PointNet#の拡張を行い、座標情報と色情報に加え時間情報を付加した点群データの学習が行えるモデルを考案し検証を行った。

1　はじめに

近年 VR や AR、MR など 3D 情報を使用した技術が発展を遂げており、また、「デジタルツイン」と呼ばれる現実空間の情報をセンサーで取得し仮想空間上に 3D で再現してシミュレーションや実験を行う取り組みも行われている。これに伴い、深層学習による 3D 情報の分類や異常検知といった技術の需要が高まっている。

3D 情報のデータ形式には、画像に対して深度情報を与えた RGB-D や立方体を積み上げて物体を表現するボクセル形式、多角形で形状を近似するポリゴン形式、３次元の点データの集合である点群形式などが存在し、データ形式によって使用できる深層学習モデルの構造は異なる。

点群形式データを扱うことが出来る深層学習モデルとしては、PointNet[1] やその発展系である PointNet++[2] などが存在する。これらの学習モデルは、座標情報 x, y, z を基幹として学習を行うモデルとなっているが、現実世界から取得された点群のデータは、座標情報に加え、色情報 R, G, B や時間情報 t を有しているため、現実世界から取得された点群データについて分類や異常検知を行うためには、それらの色情報や時間情報を加味して学習を行うことが出来るモデルが必要とされる。

著者らが以前考案した PointNet#[3] では、PointNet++を更に発展させることで色情報付き点群を学習することが可能となった一方で、時間情報が付加された点群については、扱うことが出来なかった。また、時系列 3D 情報から特徴量を抽出する手法としては、広瀬大ら [4] によりボクセル時系列データに高次局所自己相関 [5] を適応する手法が提案されている一方で、深層学習を用いて時間情報付きの点群を扱うための研究はほとんど進んでいない。

そのため、本論文では Pointnet#を拡張し、座標情報と色情報に加えて時間情報を特徴量とするモデルの考案を行い、検証として動作を撮影した時系列点群データの分類問題における精度を測定した。

*連絡先：クリスタルメソッド株式会社
E-mail: kawai@crystal-method.com

2　先行モデル

本章では、先行モデルである PointNet と PointNet++についての説明及び、拡張前の Pointnet#についての説明を行う。

2.1　PointNet

C.R.Qi et al[1] は、T-net と呼ばれる変換行列を推定する構造をモデルの中に組み込むことで、回転に不変な特徴を捉えることができる PointNet と呼ばれるアーキテクチャを提案した。また、同モデルは、モデルにプーリング層を加えることで点の格納順に左右されないグローバルな特徴量を獲得することに成功している

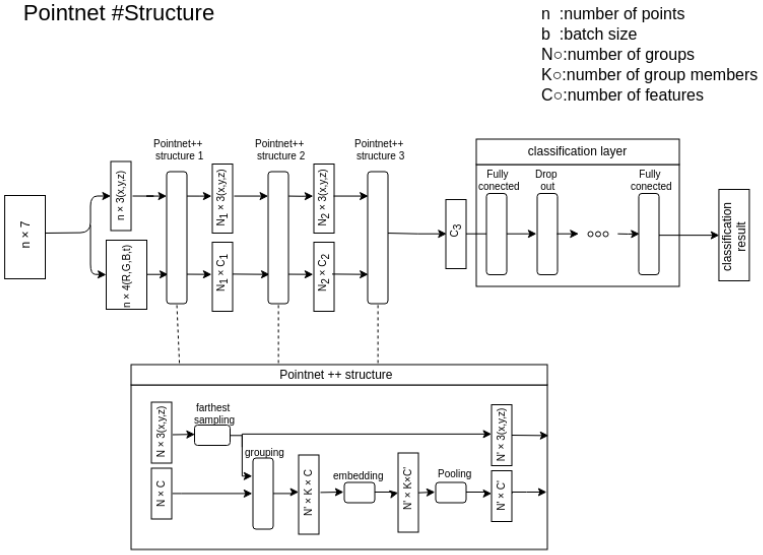

図 1: 拡張 Pointnet# 構造図[1]

一方で、このプーリング処理によって点群の局所的な情報が欠落してしまうことが知られている。

2.2 PointNet++

これに対し、C.R.Qi et al [2] は、点群を Farthest point sampling を用いて局所域に分割し、それぞれの局所域に対して PointNet のプーリング処理を適応し局所域内の特徴量を統合する処理を段階的に繰り返すことで、局所的な情報を加味しつつ、点群全体の特徴量を抽出できる PointNet++と呼ばれるアーキテクチャを提案した。局所域に対して PointNet 処理を行うレイヤーは、PointNet レイヤーと呼ばれ、PointNet レイヤーの入力サイズは、点群の保持点数を n、局所域数を k、空間軸数を d、チャンネル数を c とすると $[n, k, (d+c)]$ となる。PointNet++では、第 1 層の PointNet レイヤーへの入力チャンネル数が明示されていないが、1 層目 PointNet 層の入力に色情報や時間情報を付加することで、それらの情報を加味して学習を行える可能性がある。

2.3 PointNet#

PointNet++を更に発展させたモデルに、著者らが考案した Pointnet#[3] がある。Pointnet#は、PointNet++における Pointnet 層の入力に色情報を追加し、座標情報によるグループ化後に色情報を結合することで、色情報を加味した学習を可能とした。このことから、同様の方法で時間情報を付加することで、時間情報も加味した学習を行える可能性が高いが、PointNet#ではその可能性の言及に留まっており、検証が行われていない。

3 拡張 PointNet#モデル

上記の背景により、本研究では、PointNet#の一層目の入力に座標標情報と色情報に加え時間情報を付加することにより、時間情報を加味した学習を行えるように拡張を行う。

3.1 モデル構造

拡張後の PointNet#のモデル構造を図 1 に示す。拡張後 PointNet#は、拡張前 PointNet#と同様に 3 層の PointNet++ structure レイヤーと 1 層の classification レイヤーから構成される。

3.1.1 PointNet++レイヤー

PointNet++ structure レイヤーは、PointNet++における Sampling レイヤーと Grouping レイヤーと PointNet レイヤーを束ねたものである。

1 層目の PointNet++ structure レイヤーは、点群の保持点数を n として $(n \times 7)$ の行列を入力として受け取った後、座標情報 $(n \times 3)$ と付加情報に分割する。ここで、付加情報とは、拡張前の PointNet#では色情報のみであることに対し、拡張後 PointNet#では、色情報と時間情報 $(n \times 4)$ である。分割後、座標情報につい

[1] PointNet#[3] 図 1 を一部改変して作成

て farthest point sampling を行い、代表点を抽出する。代表点情報はサンプリング数を $N1$ として $(N1 \times 3)$ の行列として保持される。その後、代表点情報と座標情報に基づき、点のグルーピングを行う。この際、各代表点から一定距離内にある点を各グループとして定義し、グループ内の各点は、特徴量として、代表点からの相対座標に変換された座標情報とその点に対応した付加情報を保持する。そのため、この時の行列数はグループメンバー数を $K1$ として、$(N1 \times K1 \times 7)$ となる。その後、この特徴量情報について埋め込み処理により特徴量の次元を $C1$ 次元へと変換した後プーリング処理を行い、グループ内の点の特徴量の統合を行う。そのため、特徴量情報は、プーリング処理後は、$(N1 \times C1)$ の行列となり、保持していた代表点情報とともに 2 層目の PointNet++ structure レイヤーへと出力される。

2 層目の PointNet++ structure レイヤーは、1 層目の PointNet++ structure レイヤーからの出力である代表点情報（$N1 \times 3$）と特徴量情報（$N1 \times C1$）を入力として受け取る。その後、1 層目における座標情報に代表点情報を、付加情報に特徴量情報を用いて、1 層目と同じ処理を行う。拡張前 PointNet#では、座標情報と付加情報の重みを均一にするために 2 層目、3 層目の PointNet++ structure レイヤーでは、座標情報の追加を行わなかったため、拡張後の Pointnet#においても、座標情報の追加を行わなかった。

3 層目の PointNet++ structure レイヤーは、2 層目の PointNet++ structure レイヤーからの出力を入力として受け取り、2 層目と同じ処理を行う。サンプリング数を 1 として、残っているすべての点を 1 つのグループとしてみなして処理を行う。そのため、3 層目の出力は埋め込み後のチャンネル数を $C3$ として、サイズ $C3$ の特徴ベクトルとなる

3.1.2 classification レイヤー

classification レイヤーは、複数の全結合層と drop out 層からなり、3 層目の PointNet++ structure レイヤーの出力である特徴ベクトルの次元数が分類クラス数になるように変換し出力する。

3.1.3 損失関数

損失関数には、classification レイヤーの出力を softmax 関数にて活性化した結果と正解ラベルとのクロスエントロピー誤差を用いた。

4 検証実験

本章では、時間情報を扱えるように拡張した PointNet#において、座標情報と色情報に加えて時間情報を特徴量としたデータに対する検証実験を行った結果を述べる。

検証実験では、手の動作を時系列として撮影し作成した独自データセットを用いて、6 クラス分類問題の学習を行い判別精度を計測した。その際、拡張前の PointNet#において、同様のデータセットの座標情報と色情報のみを特徴量としたデータで学習を行い、時間情報を加えた場合と加えなかった場合における精度の比較を行った。

4.1 データセット

深度カメラを用いて時間による人間の手の形状の変化を撮影し、2 つのデータセットを用意した。

1 つ目は、じゃんけんに用いられる手形状 3 種（グー、チョキ、パー）のうち 1 種から他の 1 種への変化を時系列データとして撮影したものであり、このデータセットを「じゃんけんデータ」とする。「じゃんけんデータ」では、グーからチョキ、グーからパー、チョキからグー、チョキからパー、パーからグー、パーからチョキの 6 種類のデータを撮影した。撮影したデータの数は、1 クラスにつき 15 動作であり、1 動作につき時点の違う 9 つの点群を撮影した。撮影したデータの様子を図 2 に示す。図 2 の上段はチョキからパーへの変化、下段はグーからパーへの変化を時間情報が小さい順に左から並べたものである。

2 つ目は、指数えにおいて数字を表す手の形状の変化を時系列データとして撮影したものであり、このデータセットを「数データ」とする。ここで、指数えに於いて数字を表す手の形状とは、5 を数えるならば全ての指を立てた状態、4 を親指の指を閉じ他の指を立てた状態のように、数と同じ数の指を立てた手の形状を指す。本データセットでは、ある数から数を 2 つ数えるまでの動作を撮影し、1 → 2 → 3、2 → 3 → 4、3 → 2 → 1、3 → 4 → 5、4 → 3 → 2、5 → 4 → 3 の 6 種類のデーを撮影した。撮影したデータの数は、1 クラスつき 15 動作であり、1 動作につき時点の違う 9 つの点群を撮影した。撮影したデータの様子を図 3 に示す。図 3 の上段は 1 から 3 への変化、下段は 5 から 3 への変化を左から時間情報が小さい順に並べたものである。

この 2 つのデータセットそれぞれについて、次の方法によって時間情報を付加した点群を作成した。

初めに、1 動作を表す 9 つの点群から、それぞれの点群の点数が均等になるようにランダムサンプリングを行う。その後、ランダムサンプリングによって選ばれた点の座標情報と色情報 (x, y, z, r, g, b) を抽出する。

図 2: じゃんけんデータ 撮影データ

図 3: 数データ 撮影データ

この 6 つの特徴量に加えて、動作を開始してからその点群が撮影されるまでの時間の情報を時間情報として追加し、特徴量が 7 次元 (x, y, z, r, g, b, t) となる点を 10000 点の保持する点群を作成した。なお、7 次元の特徴量については、それぞれ正規化処理を行った。作成したデータの様子を図 4 に示す。図 4 の左側は、図 2 の上段のデータを用いて作成したデータの様子であり、図 4 の右側は、図 3 の下段のデータを用いて作成したデータの様子であり、表示されている各点が 7 次元 (x, y, z, r, g, b, t) の特徴量を有している。

以上の方法で、時間情報を付加した点群を作成した。各クラス撮影された動作データが 15 個であったため、その内の 10 個から学習用データを、残り 5 個から評価用データを作成した。

その際、学習用データについては、データ数が少ないため、以下の 2 つの方法を用いてデータ拡張を行った。1 つ目は、1 動作を表す 9 つの点群から 8 つのみを選択し、選択された点群達を 1 つの動作とみなし、上記と同様の方法にて時間情報を付加した点群を作成する方法である。この方法では、9 つの点群を用いた場合よりも 1 つの点群からランダムサンプリングによって選ばれる点の数が多くなり、座標情報と色情報が異なるデータが作成できる。2 つ目は、ランダムサンプリングを繰り返すことによって各点群より別の点を抽出し、座標情報と色情報が異なるデータを作成する方法である。この 2 つの方法によって、学習データについてデータ拡張を行い、各クラス 500 個の学習用データを作成した。また、評価用データはデータ拡張をせず、各クラス 5 個ずつとした。

図 4: 時間情報を付加した点群（左：じゃんけんデータ 右：数データ）

4.2 検証結果

4.1 の方法で作成したデータセット「じゃんけんデータ」と「数データ」の 2 つに対し、それぞれ 6 クラス分類問題として学習を行ったあと、評価用データにて判定精度を測定した。なお、学習にはバッチ学習を用いた。また、4.1 にて作成したデータセットから時間情報を抜いて作成した、特徴量が 6 次元 (x, y, z, r, g, b) となるデータについても拡張前の PointNet#を用いて学習を行い、時間情報がある場合とない場合の判定精度の比較を行った。

4.2.1 じゃんけんデータ 判別結果

「じゃんけんデータ」に対し、時間情報がある場合とない場合についてそれぞれ学習を行った後、評価データを分類した時のクラス別精度を表 1 に示す。表 1 の通り、時間情報ありの場合、時間情報なしの場合共に判別精度が全てのクラスに於いて 100%となった。

4.2.2 数データ 判別結果

同様に「数データ」についても時間情報がある場合とない場合についてそれぞれ学習を行った後、学習したモデルを用いて評価データの分類を行った。なお、学習にはバッチ学習を用いた。判定精度を表 2 に示す。表 2 の通り、「数データ」においても、時間情報ありの場合、時間情報なしの場合共に判別精度が全てのクラスに於いて 100%となった。

4.3 考察

4.2.1 と 4.2.2 の結果より、時間情報がある場合と時間情報が無い場合共に、どちらのデータセットにおいても全クラスの判別精度が 100%となり、時間情報がなければ判別が難しいと予想された「グーからチョキ」と「チョキからグー」のように動作が対 (逆順) となるデータについても、時系列が無い場合においても判別が行えた。この結果に関して、図 5 に評価用データとして用いられた「グーからチョキ」と「チョキからグー」のデータの一部の様子を示す。図 5 の左側が「グーからチョキ」、右側が「チョキからグー」のデータである。この 2 つのデータから見られる通り、今回のデータセットにおいては動作が対 (逆順) となるクラスのデータでも、座標情報や色情報に違いがあり、時系列情報が無くても判別が行えた可能性が高いと考察出来る。

加えて、同様の理由とデータセットの元となったデータ数が少ないことから、今回の検証で用いたデータは判別が容易であった可能性が高く、時間情報ありの場合、時間情報無しの場合共に高い精度で判別が出来たと考察出来る。

表 1: じゃんけんデータ 判別精度

	時間情報あり	時間情報なし
グーからチョキ	100%	100%
グーからパー	100%	100%
チョキからパー	100%	100%
チョキからグー	100%	100%
パーからグー	100%	100%
パーからチョキ	100%	100%
平均	100%	100%

表 2: 数データ 判別精度

	時間情報あり	時間情報なし
1 → 2 → 3	100%	100%
2 → 3 → 4	100%	100%
3 → 4 → 5	100%	100%
5 → 4 → 3	100%	100%
4 → 3 → 2	100%	100%
3 → 2 → 1	100%	100%
平均	100%	100%

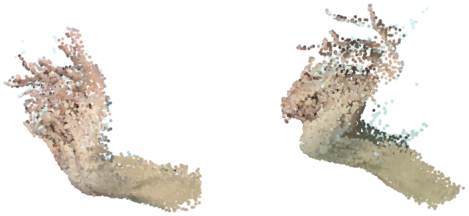

図 5: じゃんけんデータ 評価用データ

ここで、4.2 の検証では、時間情報ありの場合、時間情報無しの場合に於いて精度の差がなく、付加した時間情報が拡張後 Pointnet#において判定に用いられているかを考察することが難しかったため、次節に述べる追加検証を行った。

4.4 追加検証

追加検証では、付加した時間情報が拡張後 Pointnet#において判定に用いられているかを検証する。そのため、4.1 で述べた撮影データから次項に示す方法で、時間情報のみが異なるデータセットを作成し、4.2 の検証と同様に 6 クラス分類問題の学習を行う。

4.4.1 データセット

「じゃんけんデータ」における「グーからチョキ」と「チョキからグー」のように、「じゃんけんデータ」と「数データ」のそれぞれのクラスには、データセット内に対となる（逆順の動作を行っている）クラスが存在する。そのため、あるクラスの動作から 4.1 の手順にて時間情報を付加した点群を作成する際に、動作を開始してからその点群が撮影されるまでの時間の情報ではなく、その点群が撮影されてから動作が終了するま

での時間の情報を付加することで、対（逆順）の動作における時間情報を付加したデータを作成することが出来る。

これにより、それぞれのクラスの動作から時間情報を付加した点群を作成する際に、時間情報として、動作を開始してからその点群が撮影されるまでの時間の情報を付加した点群とその点群が撮影されてから動作が終了するまでの時間の情報を付加した点群の 2 種類を作成することで、各クラスと対となる（逆の動作を行っている) クラスにおいて、座標情報と色情報が同じで時間情報だけが異なるデータを作成することが出来る。

以上の方法で、「じゃんけんデータ」と「数データ」の 2 つのデータセットについて、座標情報と色情報が同じで時間情報だけが異なるデータを作成する。各クラスで撮影された動作データが 15 個であるため、対となるクラスで作成した動作データを含めると、各クラス 30 個の動作データが存在する。その内 20 個から学習用データを、残り 10 個を評価用データを作成した。その際、そのクラスとして撮影された動作データと対となるクラスで作成された動作データがそれぞれ同じ数だけ学習用データと評価用データに割り振られれるようにした。

また、4.2 と同様に学習データに対しデータ拡張を行い、各クラス 600 個の学習用データを作成した。また、評価用データはデータ拡張をせず、各クラス 10 個ずつとした。

4.4.2 検証結果

4.4.1 にて作成した 2 つのデータセットについて、拡張後 PointNet にてそれぞれ 6 クラス分類として学習を行い、判定精度を測定した。また、この 2 つのデータセットでは時間情報を用いなければ判別が行えないことを示すために、4.4.1 にて作成したデータセットから時間情報を抜いた、特徴量が 6 次元 (x, y, z, r, g, b) となるデータについても拡張前の PointNet#を用いて学習を行い、判定精度を測定した。表 3 に「じゃんけんデータ」のクラスごとの判別精度を、表 4 に「数データ」のクラスごとの判別精度を示す。表 3、4 より、時間情報がある場合は、それぞれのデータセットの平均判別精度が 100%となり、時間情報がない場合は、それぞれ 50%となった。

4.4.3 考察

4.4.2 の時間情報が無い場合の判別精度より、2 つのデータセットそれぞれの平均精度が 50 %であることが読み取れ、追加検証で用いたデータセットは時間情報がなければ判別不可能なデータであったと云える。また、時間情報ありの場合の判別結果より、双方のデータセットにおいて全ての評価用データを正しく判断できていると云え、拡張後 PointNet#において、時間情報が加味された判別が行えていると考察出来る。

表 3: じゃんけんデータ 判別精度

	時間情報あり	時間情報なし
グーからチョキ	100%	10%
グーからパー	100%	0%
チョキからパー	100%	100%
チョキからグー	100%	90%
パーからグー	100%	100%
パーからチョキ	100%	0%
平均	100%	50%

表 4: 数データ 判別精度

	時間情報あり	時間情報なし
$1 \rightarrow 2 \rightarrow 3$	100%	0%
$2 \rightarrow 3 \rightarrow 4$	100%	100%
$3 \rightarrow 4 \rightarrow 5$	100%	0%
$5 \rightarrow 4 \rightarrow 3$	100%	100%
$4 \rightarrow 3 \rightarrow 2$	100%	0%
$3 \rightarrow 2 \rightarrow 1$	100%	100%
平均	100%	50%

5 まとめと今後の展望

本論文では、近年の「VR 技術」や「MR 技術」、「デジタルツイン」など、現実空間の情報を 3D 情報にて再現する技術の台頭に伴う、3D 情報を用いた分類技術や異常検知技術の需要の高まりから、座標情報と色情報に加えて時間情報を特徴量として学習が行えるように PointNet#の拡張を行い、検証実験を行った。検証実験では、動作を撮影した 2 つのデータセットを用いて、時間情報を特徴量に加えた場合と加えない場合における判別精度を比較した。検証の結果、どちらの場合においても判別精度が 100%となった。そのため、時間情報が特徴量として判定に用いられているかを確かめる追加検証を行い、拡張後の PointNet#では、時間情報が特徴量として正しく判別に用いられていることが明らかになった

今回の検証では、色情報と時間情報が特徴量として付加されている公開データセットが無いことや点群の撮影コストの高いことから、検証に用いたデータセットが小規模かつ独自のものとなってしまい、判別精度

が高いものとなった可能性が大きい。そのため、大規模なデータセットを利用しての判別精度の検証を今後の課題としたい。

参考文献

[1] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In CVPR(2017).

[2] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In NIPS(2017)

[3] 河合継, 黒田唯矢 , 眞嶋啓介:PointNet#: 色情報を付与した 3 次元点群の分類, 研究報告情報基礎とアクセス技術 (IFAT),2019-IFAT-135,pp.1-6(2019)

[4] 広瀬大, 森裕紀, 浅田稔: 4 次元データに対する高次局所自己相関特徴を用いた 3 次元動画モーション認識，MIRU2012 第 15 回画像の認識・理解シンポジウム論文集，Vol.DVD-ROM, DS-22(2012).

[5] N.Otsu and T.Kurita，“A new scheme for practical flexible and intelligent vision systems ” Proc.IAPR Workshop on Computer Vision，pp.431 − 435(1988).