B-05

畳み込みニューラルネットワークによる 3 次元空間の姿勢推定

3D Pose Estimation via Convolutional Neural Network

松尾 星吾 †　廣田 敦士 †　Sai Dinesh Dacharaju ‡　岡 夏樹 †
Seigo Matsuo　Atsushi Hirota　Natsuki Oka

1. はじめに

人体の姿勢や運動情報を獲得する技術は、スポーツ競技の運動動作の解析、ジェスチャなどの動作入力によるインターフェース、CG アニメーションや映画作品など、様々な分野で応用されている。一般に、人体の姿勢情報の獲得には、モーションキャプチャが用いられている。モーションキャプチャには、体に反射マーカを取り付け、それを複数のカメラで撮影することで体の部位を測定する方法や、加速度計や角速度計を体に装着し計測されたデータから関節の動きを計算する方法、磁気センサを用いる方法などがある。また、RGB カメラと深度センサなど複数のセンサの情報から、体の部位を測定する手法もある[5]。これらの方法は、測定するための空間や特殊な機器を用意する必要があり、動作や環境が制限される。体にマーカを取り付けることが困難な環境や運動、複数のカメラを設置することができない状況ではこれらの方法を用いることはできない。一般に普及している単眼カメラから姿勢推定を行うことができれば、上述のような場合や、既存の画像、動画データの解析など、より幅広い分野での応用が可能になる。

本論文では、単眼カメラで撮影された画像からディープニューラルネットワークを用いて３次元空間での人体の姿勢を推定する手法を提案する。

2．関連研究

従来の機械学習の手法の場合、特徴量やパーツの検出器を設計し、関節間の相互作用などの人体のモデル化を行う必要があった。しかし、ディープニューラルネットワーク（DNN：Deep Neural Network）では特徴抽出に当たる部分を学習することできるため、特徴量の設計が難しい問題を解決することができる可能性がある。

２次元での人体の姿勢推定に関する研究には次のようなものがある。Toshev らは画像からの姿勢推定を回帰問題として定式化し、複数の DNN を直列につなげることで推定を行う手法を提案している[1]。この手法では、まず関節の座標のおおまかな位置の推定を行う DNN に画像を入力する。推定された座標を基準にその周囲の画像を元画像から切り取り、次の DNN に入力している。これらのネットワークには畳み込みニューラルネットワーク（CNN：Convolutional Neural Network）が用いられている。CNN は物体認識において成果をあげていたが[4]、姿勢推定のような回帰問題に対しても有効であることが示された。

Wei らは、各関節の局所的な特徴量とそれらの関係を学習することで、姿勢を推定する手法を提案した[2]。CNN を用いて構成されており、ネットワークは２段階の処理に分けられる。１段階目のネットワークは、入力された画像に対して関節が各座標にあるもっともらしさを出力するものである。この段階で関節ごとの特徴量が学習される。２段階目のネットワークでは、１段階目のネットワークで推定された各関節の座標ごとのもっともらしさと元画像が入力となり、それらの相互関係から関節の座標を推定する。体の部位によっては局所的な特徴だけではうまく推定することができず、他の部位の特徴量との相互的な関係を学習することで推測を行っている。

以上のように、CNN を用いて 2 次元での姿勢推定を行う手法が提案されてきた。本研究では、3 次元姿勢のデータセットを用意し、CNN を用いて 3 次元での姿勢推定を行う。

3. 畳み込みニューラルネットワーク

本節では、本研究で用いた畳み込みニューラルネットワークについて述べる。

3.1　ニューラルネットワーク

ニューラルネットワークとは、神経細胞を模したネットワークモデルである。細胞１つを模したユニットを層状に並べ、隣接する層ごとに互いに結合された構造を持つ。入力された情報の伝播が入力から出力へ１方向にのみ行われるものを順伝播型ネットワークと呼ぶ。ユニットは複数の入力を受け取り、それらとバイアスの総和を活性化関数の入力とする。活性化関数にはロジスティック関数や正規化線形関数などの非線形関数が用いられ、活性化関数の出力がそのユニットの出力となる。ユニットから出力され、次の層のユニットに伝わるデータには、それぞれ対応された重みがかけられる。

順伝播ネットワークは、誤差逆伝播法によって望む出力になるように重みを更新させていくことで学習を行う。学習の際は、入力に対するネットワークの出力と望む出力との差を誤差として定義し、これを小さくする。勾配降下法によって各層の重みは誤差がより小さくなる値に少しずつ変化させられる。大きなネットワークの学習はデータ 1 つ 1 つに対して行うと計算コストが非常に多くかかるため、通常は複数のデータをひとまとめにしたミニバッチと呼ばれる単位で行われる。ミニバッチに対する全データに対する誤差を計算し、その勾配の方向に重みを修正する。ミニバッチのサイズを大きくすると学習が早く進むが、バッチ間の誤差のばらつきが少なくなるため収束が遅くなるなどの影響がある[3]。

ニューラルネットにおいて、中間層を複数重ねた深い構造を持つものをディープニューラルネットワークと呼ぶ。各層の段階的な処理により、入力信号から複雑な特徴を抽出することが可能になる。

3.2　畳み込みニューラルネットワーク

畳み込みニューラルネットワークとは、主に画像処理の分野で応用される順伝播ネットワークの構造の一種である。脳の視覚野における神経細胞の受容野をニューラルネット

† 京都工芸繊維大学, Kyoto Institute of Technology

‡ Indian Institute of Technology, Guwahati

で実装している。畳み込みニューラルネットでは、畳込みとプーリングの２種類の演算を行う層が用いられる。

畳み込みとは、入力される画像サイズより小さいサイズの重みフィルタと、入力画像のある部分との積和計算のことである。畳込みには、画像の濃淡パターンがどこにあるのかを抽出するような働きがあり、学習によって目的に特化した特徴抽出器が形成される。画像にフィルタを重ねたとき、フィルタが画像の外にはみ出るような場合には元画像に未定の値を定める処理が行われ、一般的にはゼロパディングと呼ばれる 0 を設定する処理が用いられる。フィルタを適用する位置の間隔をストライドと呼び、ストライドの値を s とすると出力される画像のサイズは $1/s$ 倍になる。

プーリング層は畳み込み層の後に設置され、畳み込み層の出力をダウンサンプリングすることで対象となる特徴量の位置が変化してもプーリング層の出力を不変にする働きをしている。

一般に畳み込み層とプーリング層のペアを複数回繰り返した後、ユニット同士が全て結合している層を配置する。畳み込み層を複数回繰り返した後、プーリング層を配置する場合もある。

4. 実験

4.1 予備実験

予備実験として、CNN を用いて２次元での姿勢推定を行った。ネットワークの構成は、畳み込み層とプーリング層のペアを 2 層、全結合層を 2 層、入力は画像を 28×28 にリサイズしたカラー画像、出力は 20 関節×2 座標である。活性化関数には全て ReLU 関数を用いた。図 4.1 に実験結果を示す。赤色の×印がネットワークの出力座標、青色の◯印が正解座標である。

図 4.1 2 次元での姿勢推定の結果

1. ジェスチャや音声認識によって操作を可能にするインターフェース。RGB カメラと深度センサを搭載しており体の部位ごと位置を 3 次元で測定することができる。

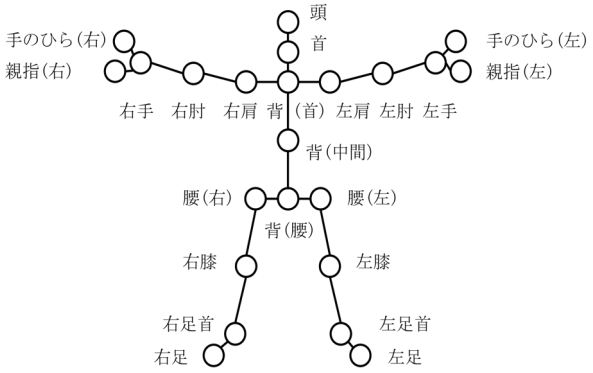

図 4.2 測定する 25 関節の位置

4.2 データセット

本実験用に Microsoft 社の Kinect[1] を用いて独自にデータセットを作成した。Kinect は、ジェスチャや音声などによる操作を可能にするインターフェースであり、RGB カメラ、音声マイク、深度センサなどの装置を搭載している。内蔵するプロセッサにより、センサの範囲内にとらえた人の顔や動きなどを認識することが可能である。実験参加者は男性５名、女性２名で、Kinect から２メートルほど離れた地点で自由にポーズを取ってもらい関節点を測定した。

実験に用いる人体モデルは、図 4.2 のように頭、指先、関節などの２５の点から構成され、１つの要素は（ｘ,ｙ,ｚ）の３つの座標を持つ。１つの姿勢は２５関節×３座標のパラメータによって表現される。

4.3 ネットワークの構成

図 4.3 に提案するネットワークの構造を示す。

入力はサイズが 192×108 のカラー画像を用いる。画素毎に RGB の値があり入力されるデータ数は 192×108×3 となる。回帰分析であるため、出力層に用いる活性化関数は恒等写像としている。上述の通り、本実験で扱う人体モデルは 25 個の点から構成され、それぞれ 3 次元の座標デ

図 4.3 本実験に用いたネットワークの構造

ータを持つため、出力のサイズは 25×3 となる。学習によってパラメータが更新されるのは、畳み込み層と全結合層のみである。

4.4 結果

図 4.4 に全結合層 1,2 の活性化関数に ReLU 関数を用いたネットワークの出力、図 4.5 に全結合層 1,2 の活性化関数にハイパボリックタンジェント関数を用いたネットワークの出力をそれぞれ示す。

図 4.4 の ReLU 関数を用いた場合は人型を出力しているが、別の画像を入力にしてもこの姿勢になるため、過学習をしておりうまく学習が行えていなかった。このネットワークでは、学習初期の段階の誤差が非常に大きくなってしまう。また、望む座標の出力は負の値も取り得ることから、ハイパボリックタンジェント関数を用いることで誤差が大きくなることを防ぎ、負の値も取る活性化関数を用いることで学習が改善されると考え、全結合層の ReLU 関数をハイパボリックタンジェント関数に変更して学習を行った。結果は図 4.5 に示す通り、それぞれの関節位置の推定がうまくできておらず、人の形になっていない。

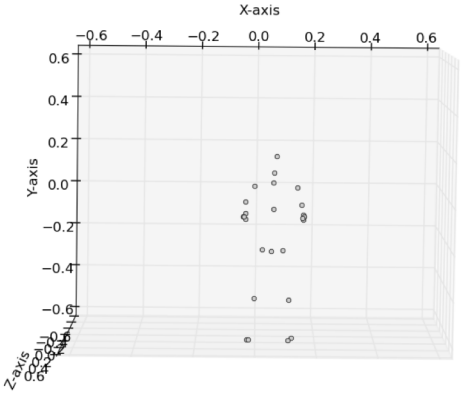

図 4.4 ReLU 関数を用いたネットワークの出力結果

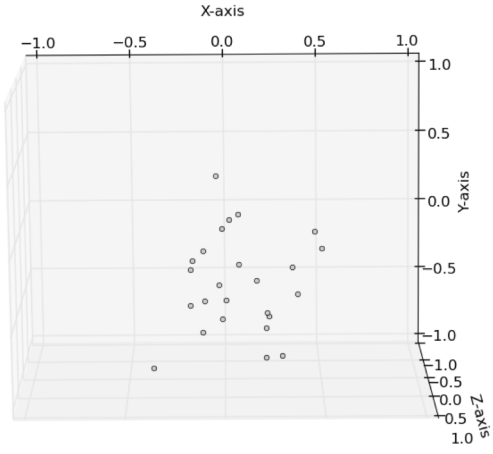

図 4.5 ハイパボリックタンジェント関数を用いたネットワークの出力結果

5. 考察

予備実験として行った 2 次元での姿勢推定の結果は、正解座標との誤差はあるが過学習にはなっておらず、入力画像と同じように腕を上げた人型が出力されているのが確認できる。しかし、同様の手順で行った 3 次元の姿勢推定は有用な結果が見られなかった。

学習がうまく行われなかった原因として考えられるのは、ネットワークの性能の限界である。Toshev らの研究では、2 次元での座標の推定に複数の DNN を用いているが、本研究では一つのネットワークで 3 次元の座標を推定しようとしている。3 次元というより複雑な問題を 1 つのネットワークを用いて解決しようとしているため、今回の手法にさらなる改善が必要である。

また、別の原因としてあげられるのは、学習に用いたデータセットの偏りである。データセット作成の実験参加者には、できるだけ体を大きく動かし止まらないように指示して自由にポーズを取ってもらったが、ポーズの間に入る直立の姿勢も撮影しているため、それがデータを偏らせたと考えられる。図 4.4 のように直立した姿勢が出力されるのは、データの偏りによって訓練誤差が小さくなるように直立に近い姿勢ばかりが学習され、汎化性能が過度に低くなっているのだとすれば、より偏りが無く多様な姿勢のデータが必要である。

6. おわりに

本研究では、DNN を用いて 2 次元画像からの 3 次元姿勢の推定を試みたが、現時点では有用な結果を示すことができていない。ネットワークの構造、学習に用いるデータセットに対し、さらなる改善が必要である。

参考文献

[1] Alexander Toshev, Christian Szegedy. DeepPose: Human Pose Estimation via Deep Neural Networks. In *Conference on Computer Vision and Pattern Recognition*, 2014.

[2] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, Yaser Sheikh. Convolutional Pose Machines. In *Conference on Computer Vision and Pattern Recognition*, 2016.

[3] Geoffrey E. Hinton. A Practical Guide to Training Restricted Boltzmann Machines. In *Neural Networks: Tricks of the Trade*, pp 599-619, 2012.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, Vol. 25, pp. 1106−1114, 2012.

[5] Jamie Shotton, Andrew Fitzgibbon, Andrew Blake, Alex Kipman, Mark Finocchio, Richard Moore, Toby Sharp. Real-time human pose recognition in parts from single depth images. In *Conference on Computer Vision and Pattern Recognition*, 2011.