Vol.2020-EC-55 No.23
2020/3/19

GANによる教師なし3次元姿勢推定の精度向上に向けての一検討

田中 哉汰[†1]　藤井 叙人[1,†1]　片寄 晴弘[1,†1]

概要：
単一の画像から人間の姿勢推定を行うタスクは現在研究が進められている分野であり，特に 3 次元座標の推定は様々な分野での応用が期待されている．Kudo らは二次元姿勢推定器によって得られた 2 次元姿勢情報を入力とし，そこから推定される奥行き情報を付加した 3 次元情報を他視点から評価し，生じるズレを最小化するという問題を GAN の枠組みによってモデル化した．本研究では Kudo らの研究を基礎として，Generator と Discriminator のネットワーク構成の変更，それらを起点とした勾配消失問題への対処を実施し，精度向上の検討をおこなった．ネットワーク構成をより複雑なものに変更し，学習に条件を設けることで従来よりも約 12 %高い精度を得ることができた．

1. はじめに

人物の姿勢推定はさまざまな領域でニーズがあり，2010 年代にはゲーム用ジェスチャ入力用デバイスとして開発された Kinect が大きな注目を集めた．その後，2017 年に発表された深層学習による姿勢推定 [1] が提案されて以降，深層学習による画像（映像）からの 2 次元，さらには，3 次元の姿勢推定に関する研究が積極的に実施されている [2], [3].

機械学習による 3 次元の姿勢推定手法は大きく，画像情報から直接 3 次元座標を推論する手法 [2] と一度 2 次元の姿勢推定を行った後，その関節座標を用いて 3 次元の関節角度を推論する手法に分けられる．後者の代表的な研究としては，Julieta ら [3]，Kudo ら [4] の取り組みが挙げられる．julieta ら [3] は，2 次元座標から 3 次元の姿勢推定を Dropout，Batch Normalization，ReLU，Redisual network から成るネットワークにより実現しており，他の提案と比べて高い精度が得られたとの報告がなされている．一方，Kudo ら [4] は，GAN(Generative Adversarial Networks) を用いて x, y 座標から z 座標を推定するユニークな手法を提案している．

本研究では，Kudo ら [4] をベースに，勾配消失問題への対策を導入し，精度向上に向けた検討を実施する．

[1] 情報処理学会
IPSJ, Chiyoda, Tokyo 101–0062, Japan
[†1] 現在，関西学院大学
Presently with Kwansei Uniersity

2. GAN による奥行き座標の推定モデル

Kudo ら [4] は，二次元姿勢推定器によって得られた 2 次元姿勢情報を入力とし，そこから推定される奥行き情報を付加した 3 次元情報を他視点から評価して生じるズレを最小化するという問題を GAN の枠組みによってモデル化した（図 1 参照）.

ジェネレータの入力データとして 2 次元の座標 $p = [p_1 ... p_N]^T = \begin{bmatrix} x_1 & ... & x_N \\ y_1 & ... & y_N \end{bmatrix}^T$ を用いる. 出力は推定した奥行き座標 $z = [z_1 ... z_N]^T$ とする．その後，股下の座標を基準として角度 θ (-π , π) 回転させたもの ρ は次のように表される.

$$\rho = f(p, z; \theta) = f(p, G(p); \theta) = p \begin{bmatrix} cos\theta & 0 \\ 0 & 1 \end{bmatrix} + z \begin{bmatrix} sin\theta & 0 \end{bmatrix} \quad (1)$$

モデル全体を以下のミニマックス問題に定式化する.

$$V(G, D) = \mathbb{E}_p[logD(p)] + \mathbb{E}_{p,\theta}[log(1 - D(\rho))] \quad (2)$$

このモデルは多視点からの形を想像して評価するというヒトの 3 次元情報のプロセスをシミュレートするもので，原理的に大きな可能性を持つと考える．その一方で，Kudo らのネットワーク構成は比較的シンプルなものだということもあって，他に提案された手法との比較では必ずしも十分な精度が得られていない．そこで，本研究では，図 1 で

ⓒ 2020 Information Processing Society of Japan

の Generator と Discriminator のネットワーク構成の変更，それらを起点とした勾配消失問題への対処を実施し，精度向上の検討を行う．

3. Generator と Discriminator のネットワーク構成の変更

Kudo らのネットワーク構成はシンプルであり，また Generator と Discriminator に同じネットワークを適応しているため精度が低いと考えられる．そこで，2 次元座標から 3 次元姿勢を推定するネットワークにおいて高い精度が得られている Julieta らのネットワークを基礎として Generator と Discriminator を変更する．julieta らのネットワーク構成は Dropout，Batch Normalization，ReLU，Redisual network から成る構造になっており，本研究ではこれらを取り入れたネットワークを構築した（図 2，3）．

Generator は 4 層の隠れ層において Dropout，Batch Normalization，ReLU，Redisual network を適用し，入力層，出力層は線形層として構築している．Discriminator は勾配消失問題を避けるため，Generator よりも単純なネットワーク構成にしており，Generator を基礎として，層数，ノード数を減らすことでネットワークを構築している．

4. 勾配消失問題への対策

勾配消失問題とは Discriminator の予測精度が上昇し，偽物と本物を見分ける能力が高くなることで精度向上が見込めなくなることであり，学習においては Discriminator の予測精度を高くしすぎないための工夫が必要となる．

本研究では以下の 3 点により，Discriminator と Generator の学習におけるバランスを取っている．

・Generator の学習頻度を Discriminator の二倍
・Discriminator の予測精度が 85 %を超えた時に学習を一時停止
・Discriminator のネットワークを単純な構成

勾配消失問題を引き起こさないために，GAN における学習では Generator の損失関数の出力を 1.4 程度に保った状態で学習を進めるとよいとされている．本研究では学習頻度と学習の一時停止は損失関数の出力が最も 1.4 に近い，2 倍，85 %に決定した．また，Discriminator のネットワークを Generator と比べて単純な構成にすることで，Discriminator の予測精度を減少させている．上記工夫を取り入れた結果，本手法における Generator の損失関数の出力が 1.5 程度で安定しており，学習が進んでいることがわかる（図 4 参照）．

5. 性能評価

本研究の学習データ，評価データには Human3.6M[5] から約 40 万 (421744) の 2 次元骨格情報，約 10 万 (105436) の 3 次元骨格データを使用しており，二つのデータセットに重複はない．このデータセットは俳優が日常の 15 シーン（食事，散歩，喫煙等）を模したデータで成り立っており，2 次元，3 次元の真値だけでなく，カメラパラメータなどの情報も提供されている．

表 1 に 2 次元関節座標から 3 次元関節座標を生成した結果を示す．また，生成した画像を図 5 に示す．図 5 は左から本研究手法における推論結果，元画像，Kudo らの手法における推論結果を示している．

Kudo らの手法と比較すると「Purchases」，「Sitting」，「Phoning」を除く 15 のシーン中 12 のシーンにおいて精度が向上しており，平均の精度は 11.8 %精度向上した．これは各関節に対して 13.8mm 精度向上したことを示しており，全体としては 234.6mm 精度向上した．従来手法よりも精度が下がった「Purchases」，「Sitting」クラスにはしゃがみ姿勢のデータ数が多いという特徴があり，原因の一つとして考えられる．

6. おわりに

本研究では Kudo らの GAN による 3 次元推定からネットワーク構成を変更することでより高い精度を示した．Generator と Discriminator のネットワークを別々に用意し，Discriminator の学習が進みすぎないように Discriminator の学習上限を設定，Generator の学習頻度に調整を施すことによって，より高い精度で 3 次元座標推定を行えることを示した．

今後の課題として以下二点が挙げられる．どちらの課題も映像を取り扱う場合に限られるが，時系列データとしての特性を生かし LSTM を用いて学習をすることでより高い精度で推測を行う．また，平面推定と組み合わせることで推定する 3 次元座標をモデル座標系からワールド座標系に拡張する．

参考文献

[1] Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh, Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. CVPR(2017)
[2] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, Pascal Fua, Structured prediction of 3d human pose with deep neural networks. BMVC(2016)
[3] Julieta Martinez, Rayat Hossain, Javier Romero, James J. Little, A simple yet effective baseline for 3d human pose estiomation. ICCV(2017)
[4] Yasunori Kudo, Keisuke Ogaki, Yusuke Matsui, Yuri Odagiri, Unsupervised Adversarial Learning of 3D Human Pose from 2D Joint Locations. CVPR(2016)
[5] Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C., Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. TPAMI(2014)

ⓒ 2020 Information Processing Society of Japan

付録

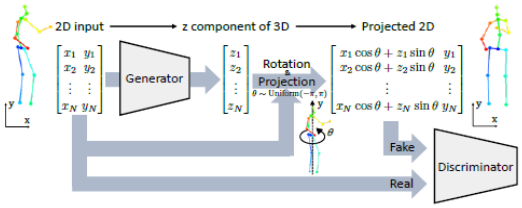

図 1　Kudo(2018) の奥行き座標の推定モデル

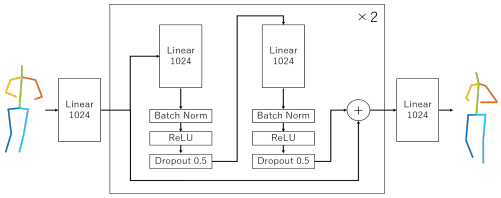

図 2　変更後のネットワーク (Generator)

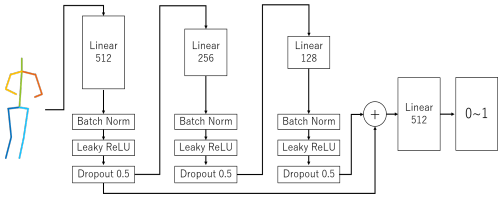

図 3　変更後のネットワーク (Discriminator)

図 5　生成結果比較

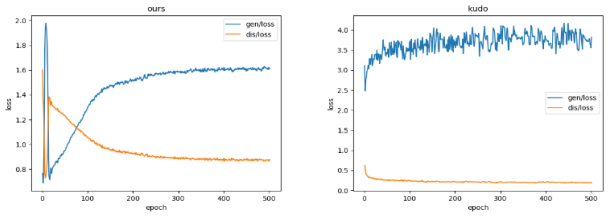

図 4　損失関数比較

表 1　推定誤差 (mm)

Method	Directions	Discussion	Eating	Greeting	Phoning	Photo	Posing	Purchases
ours	**93.1**	**102.7**	**122.4**	**104.6**	119.1	**132.5**	**88.2**	159
Kudo[4]	125	137.9	107.2	130.8	**115.1**	127.3	147.7	**128.7**
Method	Sitting	SittingD	Smoking	Waiting	WalkDog	Walking	WalkT	average
ours	164.9	**137.7**	**114.4**	**112.2**	**121.9**	**88.3**	**96.2**	**117.1**
Kudo[4]	**134.7**	139.8	114.5	147.1	130.8	125.6	151.1	130.9

ⓒ 2020 Information Processing Society of Japan