機械学習による室内における日常動作解析のための合成動画像データセット構築に向けて

儀井葉那　(指導教員：小口正人)

1　はじめに

ディープニューラルネットワーク (DNN) により動画像から人間の行動を分析することが可能になり, 一般家庭における老人や子供の見守りなどへの応用が期待されている [1]. DNN を用いた学習では, 大量かつバリエーション豊富なラベル付き学習データが必要となるが, そのようなデータセットは現在存在しておらず, また, データセットを現実の画像で作成するには多大なコストを要する. 物体検出などの一部のコンピュータビジョンタスクにおいては, 現実のデータ収集の問題に対応するために CG により合成画像を生成して学習データとする試みが行われているが, 動画像分類のための合成データ生成に関してはほとんど知られていない.

本研究では, 人間の室内行動解析ためのデータセットの構築, および現実の動作解析のための合成動画像生成の方法を確立することを目指し, Unity を用いて合成動画像データセットを作成・評価した.

2　作成した合成動画像データセット

合成動画像データセットを作成するため, Unity による動画像の作成, 作成した動画像内の照明条件の変更, および画像の劣化を模したノイズ・ぼかし処理 [2] を施す.

行動解析のためのデータセットの作成に Unity Technologies 社が提供するゲームエンジン Untiy[3] を使用した. 作成した動画像は, 部屋の中を人型モデルがランダムに歩き回る・立ち止まるという動作をし, ソファ前に来ると座り, 数秒後に立ち上がる動作を部屋の四隅上部から 256*256 ピクセル, 5fps で撮影したものである. 図 1 に座って立ち上がる様子を示す.

照明条件の変更を表現するため, 動画像内の 2 つの照明についてランダムに明るさを変更・移動させた. ノイズ・ぼかしはガウスノイズ・ガウスフィルタを用いて表現する. まず, ノイズ処理は式 (1) のようにモデル化した.

$$I_{noise}(x,y) = max(min(I(x,y)+\eta_{gauss}, 0), 255) \quad (1)$$

ここで $I_{noise}(x,y)$ は処理後の位置 (x,y) における画像の値, $I(x,y)$ は元の画像の位置 (x,y) の値, η_{gauss} はガウス分布に基づく値である.

図1　座って立ち上がる様子

図2　ぼかし・ノイズを施した 1 フレーム

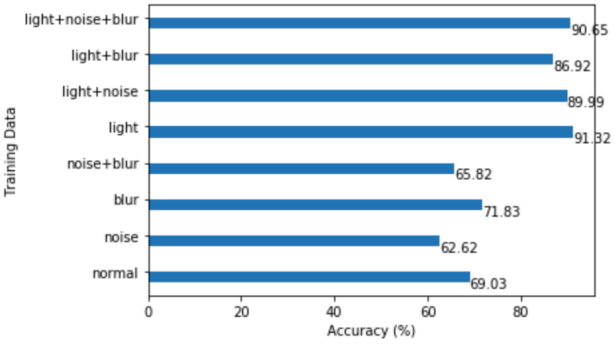

図3　作成したデータによる動作判別の精度

次に, ぼかし処理を式 (2) で表す. $I_{blur}(x,y)$ は処理後の位置 (x,y) の画像の値, $K(m,n)$ は二次元ガウス分布に基づくカーネルである.

$$I_{blur}(x,y) = \sum_m \sum_n I(x+m, y+n)K(m,n) \quad (2)$$

作成した動画像の 1 フレームにぼかし・ノイズを加えた画像を図 2 に示す.

3　評価

3.1　実験 1

本実験では, 作成した合成データで学習したモデルで合成データについての動作判別ができるかどうかを確認する. また, 照明条件のランダム化等により精度が向上することを確かめるために, それぞれの有無を変更したデータで学習を行い, テスト精度を比較する. 入力データは, 作成した 16 フレームの画像をまとめて 1 つの動画像としたものを 4000 個用意し, 学習用・検証用・テスト用に 10:3:3 に分割した. テストデータには, 照明条件のランダム化と一定の強さのノイズ・ぼかしを施した合成動画像データを用いた. 学習モデルには 18 層 ResNet3D [4] を用いて歩く・立ち止まる・座る・座っている・立ち上がるという 5 つの動作の分類を行う. 計算には GeForce GTX 980 GPU を備えたサーバ機を用いた.

学習データの条件を変化させて学習させ, テストを行

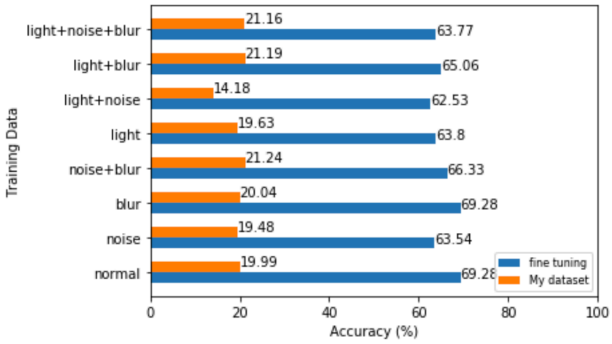

図4　STAIR-actions による動作判別の精度

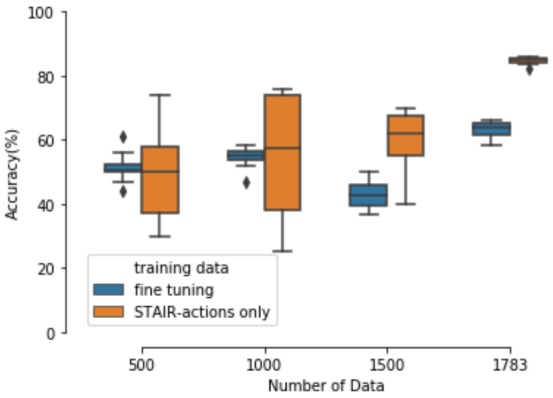

図5　データサイズと精度の比較

なった結果を図 3 に示す．図 3 から，作成した合成動画像によるテストにおいて，照明条件のランダム化を行うと精度が高くなること，ノイズ・ぼかし処理による影響はあまり見られないことがわかった．照明条件を変化させたデータで学習したモデルでは約 90%，そうでないモデルでは約 60-70% の精度で動作判別ができることがわかった．

3.2　実験 2

本実験では，実写動画像データセット STAIR-actions[5] を用いて作成したデータセットを評価する．

まず，作成したデータセットで学習したモデルをそのまま用いてテストする．次に，STAIR-actions を用いて作成したデータで学習したモデルの線形層の再学習を行う．さらに，学習データの量を変化させて，テスト精度を比較する．実験の条件は実験 1 と同じである．

1783 個のデータでの実験結果を図 4 に示す．オレンジ色の横棒は作成したデータのみで学習したモデルでの精度を，青色の横棒は作成したデータで事前学習した後に STAIR-actions でファインチューニングしたモデルでの精度を表す．図 4 より，作成した合成動画像で学習したモデルではランダムに判別するのと同等な 20% 程度の精度となり，STAIR-actions の動作をほぼ判別できていないことがわかる．しかし，STAIR-actions でモデルを学習する場合の 84.56% には及ばないが，ファインチューニングを行うと，約 60-70% の精度で判別している．このことから，作成したデータは，ある程度は現実の動作の特徴と共通する特徴を持っていることがわかる．いずれの場合も，合成データへの照明変化・ノイズ・ぼかし等による影響は見られなかった．さらに，データ量を変化させた結果を図 5 に示す．図 5 より，1500 個程度のデータを用意できる場合は STAIR-actions のみで学習したモデルの方が高精度であるが，それより少ない場合は，本データセットで事前学習したモデルでファインチューニングした場合と同等の精度であり，事前学習モデルの方が安定した学習結果になることがわかった．これらから，利用できる実写データが少ない場合には，作成したデータセットによる事前学習が有用であるとことがわかった．

4　まとめと今後の取り組み

本研究では室内における人間の行動解析のための合成動画像データセット作成を目指し，Unity で CG アニメーションをキャプチャしてラベル付き動画像データセットを作成した．動画像は 1 人の人型モデルが歩く・立ち止まる・座る・座っている・立ち上がるという 5 つの動作を行うもので，さらに，現実の動画像との差分を減らすために照明条件のランダム化，ノイズ・ぼかし処理を施した．実写データを用いた評価にて，現段階では作成したデータでは現実と CG とのギャップを埋められていないことが示されたが，ある程度は現実の動作と同等な特徴を持っていること，実写データでファインチューニングを行うことで，実写データのみで学習するのと同等な精度でより安定的に動作判別するモデルを構築できることがわかった．また現段階では，照明条件のランダム化，ノイズ・ぼかしの付与による実写データ解析への影響は確認できなかった．

今後は実写画像を用いなくても十分に実写画像を判別できるように，動作・人・背景を多様化し，データセットの拡張を行う．また，データ拡張やドメイン適応を調査および評価し，合成データによる動画像認識タスクのためのより高精度なモデルを構築することを目指す．

謝辞

この成果の一部は，JSPS 科研費 JP19H04089 および国立研究開発法人新エネルギー・産業技術総合開発機構 (NEDO) の委託業務の結果得られたものです．

参考文献

[1] Chikako Takasaki, Atsuko Takefusa, Hidemoto Nakada, and Masato Oguch. A study of action recognition using pose data toward distributed processing over edge and cloud.

[2] Alexandra Carlson, Katherine A. Skinner, Ram Vasudevan, and Matthew Johnson-Roberson. Modeling camera effects to improve visual learning from synthetic data. In *ECCV Workshops*, 2018.

[3] Unity. https://unity.com.

[4] Du Tran, Hong xiu Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2017.

[5] Yuya Yoshikawa, Jiaqing Lin, and Akikazu Takeuchi. Stair actions: A video dataset of everyday home actions. *ArXiv*, Vol. abs/1804.04326, ,