1R-02

機械学習向け生活空間動画像データセット構築の検討

礒井葉那 †　竹房あつ子 ‡　中田秀基 §　小口 正人 †

† お茶の水女子大学　‡ 国立情報学研究所　§ 産業技術総合研究所

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, 一般家庭で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 [2], 室内における人間の行動解析のためのデータセットは現状存在しない. また, そのようなデータセットを実写画像で作成するには, 多大な手間やコストを要する.

本研究では, 人間の室内行動解析のデータセット構築を目指し, 3D ゲームエンジンの Unity を用いた合成動画像データセットを試作した. 作成したデータセットでは 室内で人が歩く・立ち止まる・座る・座っている・立ち上がるの 5 つの動作をランダムに行う. また, 動画像を実写画像に似せるため, 照明条件のランダム化とノイズ・ぼかし処理を施した. 予備実験では, 作成したデータで上記の動作分類ができることを確認した.

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, センサの劣化を模したノイズ・ぼかし処理 [5] を行った.

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. 作成した動画像では, 部屋の中を人型モデルがランダムに歩き回る・立ち止まるの動作をし, ソファ前に来ると座り, 数秒後に立ち上がる動作を繰り返す. データセットでは, 各動画像を 0.2 秒ごとに 256 * 256 ピクセルの大きさでキャプチャしたものを用意した. 図 1 に座って立ち上がる様子を, 図 2 には室内の様子を表す動画像の 1 フレームをキャプチャしたものを示す.

図 1　座って立ち上がる様子

図 2　作成した動画像の 1 フレーム

照明条件の変更を表現するため, 動画像内の 2 つの照明をランダムに明るさを変更・移動させた. また, 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.

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

A Study on Development of Video Dataset in Living Space for Machine Learning
†Hana Isoi
‡Atsuko Takehusa
§Hidemoto Nakada
†Masato Oguchi
†Ochanomizu University
‡National Institute of Informatics
§National Institute of Advanced Industrial Science and Technology (AIST)

Copyright ©2020 Information Processing Society of Japan.
All Rights Reserved.

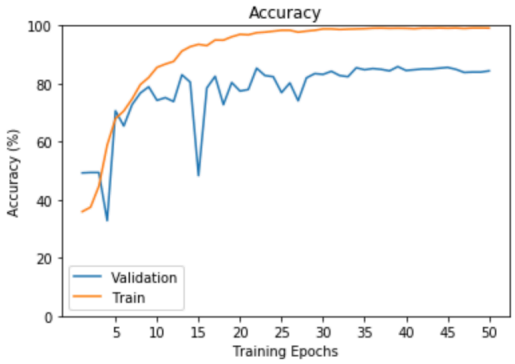

図 3　データ数 4,000, カメラ数 8 での動作識別精度

表 1　データ数とカメラ個数を変化させたテスト精度の比較

データ数 (動画像数)	4 カメラ	8 カメラ
1,000	73.91%	72.95%
2,000	81.82%	73.80%
4,000	91.19%	83.58%

に基づくカーネルである.

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

3　予備実験

実験では, 作成した合成データで動作識別ができることを確認する. また, 設置するカメラの個数を 4 個, 8 個の 2 通りに変化させて実験し, 学習精度を比較する. さらに, 照明条件のランダム化とノイズ・ぼかし処理を施し, 動作識別ができることを確認する.

実験では, 作成した 16 フレームの動画像データを 10:3:3 に分割して学習データ・検証データ・テストデータとし, それぞれ 16 フレームをまとめて 1 つの入力とし, 3D ResNet[4] を用いて Walking, Standing, Sit Down, Sitting, Stand Up の 5 クラス分類を行う. 計算には GeForce GTX 980 GPU を用いた.

照明条件のランダム化・ノイズなどを加えずにデータ数 4,000, カメラ数 8 個で 50 エポック学習させた実験結果を図 3 に示す. 図 3 から, データを加工せずにデータ数 4,000, カメラ個数 8 で学習した場合はは約 84% の精度で 5 クラスの動作分類ができることがわかった. カメラ数・データ数を変化させた場合の結果は表 1 のようになり, カメラ数が少なくデータ数が多い方が, テスト精度が高くなることがわかった.

次に, カメラ数 8, データ数 4,000 で照明条件のランダム化・ノイズ・ぼかしを施した結果を図 2 に示す. ノイズ・ぼかしを施した場合, 施さない場合と同等な精度で学習できることが示された. 照明条件のランダム化を施した場合でも, 精度の低下が見られるが, 約 75 % の精度で動作識別ができていることがわかった.

表 2　テスト精度の比較

データ	テスト精度
加工なし	83.58%
照明条件ランダム化	75.17%
ノイズ	84.51%
ぼかし	83.71%

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を目的とする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動画像を作成した. また, 現実の動画像解析に向けて, 照明条件のランダム化, ノイズ・ぼかし処理を施した. 1 人の人型モデルが歩く・立ち止まる・座る・座っている・立ち上がるの 5 つの動作をする動画像データセットを試作し, 予備実験にてそれらを判別できることを確認した.

今後は動作・人・背景のオブジェクトの多様化を行い, バリエーション豊富なデータセット作成に取り組む. また, 作成したデータセットで学習したモデルを現実のデータでテストし, その有用性を検証する.

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] Wu, D., Sharma, N. and Blumenstein, M.: Recent advances in video-based human action recognition using deep learning: A review, *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 2865–2872 (online), 10.1109/IJCNN.2017.7966210 (2017).
[2] Sun, C., Shrivastava, A., Singh, S. and Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era, *Proceedings of the IEEE international conference on computer vision*, pp. 843–852 (2017).
[3] : Unity, `https://unity.com`.
[4] Hara, K., Kataoka, H. and Satoh, Y.: Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6546–6555 (2018).

Copyright ©2020 Information Processing Society of Japan.
All Rights Reserved.