動作分類のための合成データを用いた動画像ドメイン適応

理学専攻・情報科学コース 2040631 礒井葉那

1 はじめに

近年のディープニューラルネットワークの進歩に伴う学習データ不足の問題について様々な議論が行われており, その解決策の 1 つに合成データを利用した学習がある. コンピュータにより人工的に生成される合成データには, 大量かつ多様なデータ生成が比較的容易であり, プライバシの問題が起こらないという利点がある. しかし, 合成データのみを用いて学習したネットワークでの実データ解析時にはドメインシフトが起こるため, ドメイン適応を行う学習が必要である.

本研究では, 実動画像と写実的な合成動画像からなる Ochahouse Dataset を作成する. また, Ochahouse Dataset を用いた実験から, 合成データを用いたドメイン適応を行う学習によって実データのラベルを使わずに高精度な実データの動画像分類が可能であることを示す. また, 複数の学習手法を比較し, 合成動画像ドメイン適応に効果的な学習手法について考察する.

2 合成データセット作成の検討

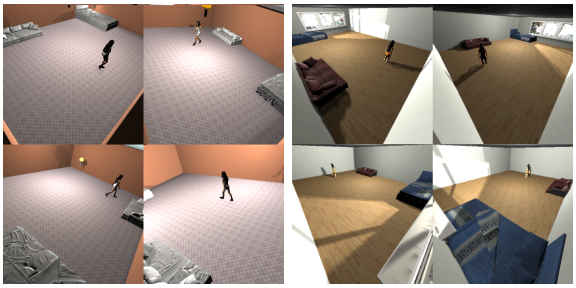
実動画像分類に適した合成動画像データを作成する方法を検討するため, 3 種類の合成動画像データを作成する. また, それらを用いて学習したモデルで既存実動画像データ STAIR-Actions の動作分類を行い精度を比較することで, 適切な合成動画像データの特徴を明らかにする.

作成した 3 種類の合成動画像データ A, B, C (図 1(a),(b),(c) に示す) はすべて, Unity ® で作成された, 部屋の中を人が動き回り, 歩く, 座る, 立ち上がるの 3 つの動作をする様子を収録したものである. これらの部屋は同じ広さ・構造で, カメラの数と位置も同様であるが, 色やテクスチャが異なっている. 合成動画像 A の部屋は簡易的な内装, 合成動画像 B の部屋はより写実的な内装であり, 合成動画像 C では合成動画像 B と同様の部屋で, 拡大・追跡カメラによって収録されている.

作成した動画像 A, B, C を用いて, 既存動画像解析モデル 3D ResNet-18 を 100 エポック学習させ, STAIR-Actions データセットの該当する 3 クラスのサブセットの動作クラスの判別を行う. 学習, 検証に用いたデータ数はそれぞれ 2500, 750 であり, テストに用いた STAIR-Action データ数は 1784 である. 各データで学習したモデルを用いた STAIR-Actions のテスト精度を表 1 に示す. 学習データ A と B の比較では, B のほうが高精度なことから, 部屋の内装がターゲットデータに近いデータの方が有効であることがわかった. また, 学習データ B,C の比較では B のほうが分類精度が高く, 人がうつる大きさを大きくすることは効果がないことがわかった.

3 Ochahouse Dataset

前章の結果をもとに, 部屋の中を 1 人の人が動き周り, walking, sitting down, sitting, standing up, lying down, lying, getting up の 7 つの動作をする実動画像 Ochahouse-Real と, 合成動画像 Ochahouse-Syn を作成した. 合成動画像 Ochahouse-Syn の作成には Unity を使用した. 各動画像は約 3 秒から 7 秒程度である. それぞれのデータの各動作クラスとデータ数を表 2に, 作成した実動画像データと合成動画像データの 1 フレームを図 2に示す.



(a) A (b) B

(c) C (d) STAIR-Actions

図1 (a)(b)(c) それぞれデータ A, B, C の 1 フレームの例. (d) 既存実動画像データ STAIR-Actions の 1 フレームの例

表1 各動画像で学習した STAIR-Actions の精度

学習データ	STAIR-Actions の分類精度 (%)
A	32.66
B	43.29
C	35.00
STAIR-Actions	84.56

4 DANN によるドメイン適応

ドメイン適応とは, ドメイン間に共通する特徴を学習させることでドメインシフトに対応する転移学習の一種の手法である. 教師なしドメイン適応の代表的な手法である DANN (Domain Adversarial Neural Networks)[1] により 3D ResNet を拡張した図 3のモデルでは, クラス分類器によりクラス分類ロス $\mathcal{L}_{class}$ を最小化する一般的な学習と, ドメイン分類器 (弁別器) でドメイン分類ロス $\mathcal{L}_{domain}$ を最小化する敵対的学習を同時に進める, 以下の最適化を行う.

$$\min \mathcal{L}_{class} + \alpha \mathcal{L}_{domain}$$

表2　Ochahouse Dataset の動作クラスと各データ数

クラス	walking	sitting down	sitting	standing up	lying down	lying	getting up
合成データ	997	747	1118	780	250	250	250
実データ	96	44	56	51	32	39	32

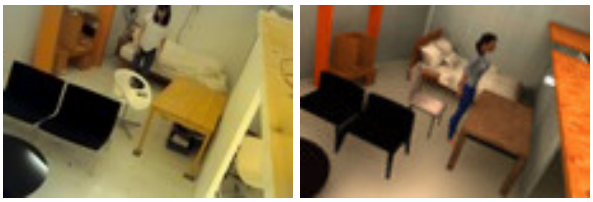


図2　実動画像データ Ochahouse-Real の 1 フレーム (左) と合成動画像データ Ochahouse-Syn の 1 フレーム (右)

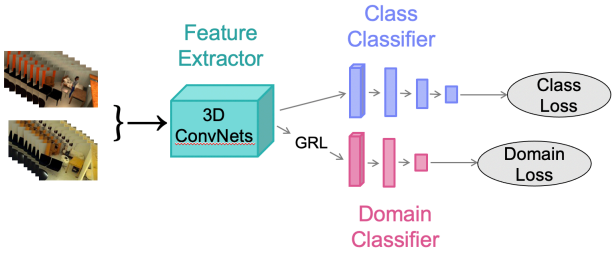
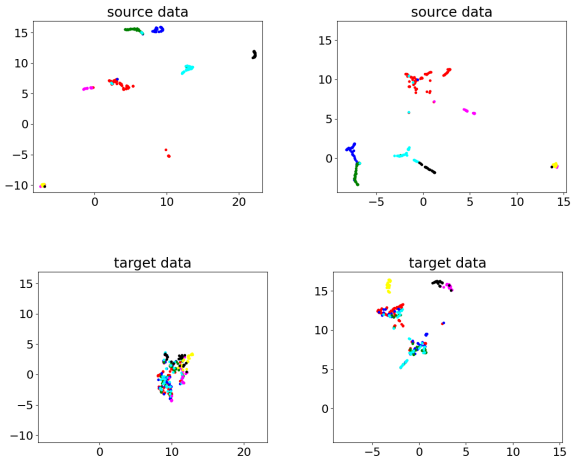


図3　3D ResNet を DANN によるドメイン適応を行うよう拡張したモデル

表3　実データ動作分類精度 (%)

ベースモデル	教師あり	教師なし	教師なしドメイン適応
3D ResNet	88.60	11.11	40.74
TSN	78.13	25.64	37.27
TRN	66.67	46.41	63.08
TA^{3}N	66.92	48.03	55.77



(a)　(b)

図4　(a) TA^{3}N での教師なし学習, (b) TA^{3}N での教師なしドメイン適応による学習の各モデルで抽出される合成動画像と実動画像の特徴量の UMAP による可視化.

5　実験

Ochahouse-Syn を用いてドメイン適応を含む複数の手法で 60 エポック学習し, Ochahouse-Real の動作分類を行う. 動画像の特徴抽出には 3D ResNet, TSN, TRN のそれぞれを, DANN によるドメイン適応を行うよう拡張したモデルと, 既存動画像ドメイン適応モデル TA^{3}N[2] を使用した. 計算には 1 台の Tesla V100 PCIe 32GB を用いた.

実験結果は表 3のようになり, 各モデルにおいて教師なし学習での精度は教師あり学習での精度より低く, ドメインシフトが起こっていること, 教師なしドメイン適応により精度が向上していることがわかった. また, TRN, TA^{3}N をベースとするモデルではドメイン適応によって教師あり学習と同程度の精度まで動作分類精度が改善しており, これらに含まれる時間関係推論を行うモジュールがドメインシフトの解消に有効であることがわかった.

また, TA^{3}N による教師なし学習, TA^{3}N による教師なしドメイン適応で抽出される Ochahouse-Syn, Ochahouse-Real それぞれの特徴量ベクトルを UMAP で 2 次元に圧縮しプロットしたものを図 4に示す. 各図の source data は Ochahouse-Syn の動画像特徴表現ベクトルを, target data は Ochahouse-Real の動画像特徴表現ベクトルを表す. 図 4(a) では, 各モデルで抽出された動画像特徴表現ベクトルは source data と target data で分布の形状が異なっていることからドメインシフトが起こっていること, (b) ではドメイン適応によって分布の形状が近づいたが改善の余地があることがわかる.

6　まとめと今後の取り組み

本研究では合成動画像を活用した教師なし学習による高精度な動作認識の実現を目指して, 合成動画像データセットを作成し, 複数のドメイン適応を用いた学習手法を比較した. 実験から, 作成した合成データは我々人間の目で見て写実的であるが実データ解析時にはドメインシフトが起こること, DANN によるドメイン適応と TRN による時間関係推論によって教師あり学習と同程度に高精度な分類ができるようになることがわかった.

今後は, これらの結果をもとに効率的な動画像ドメイン適応手法を提案し高精度な教師なし動作分類を実現することで, コストやプライバシの問題などでラベル付き実データの用意が困難であるという課題の解決を図る.

参考文献

[1] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P.,Larochelle, H., Laviolette, F., Marchand, M. and Lempit-sky, V.: Domain-Adversarial Training of Neural Networks(2016)

[2] Chen, M.-H., Kira, Z., AlRegib, G., Yoo, J.,Chen, R. and Zheng, J.: Temporal Attentive Alignment for Large-Scale Video Domain Adap-tation, ICCV2019