6U-02

合成データを用いた教師なしドメイン適応による室内動作認識手法の比較

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. 合成データには生成が比較的容易であるという利点があるが, 合成データを用いて学習したモデルには, 実データ解析時にデータの特徴量分布の違いから解析精度が低下するドメインシフトが起こるという課題がある.

ドメイン適応とは, ドメインシフトに対応するための手法であり, 合成データで学習された分類器を実データに用いる場合に必要とされることが知られている. ドメイン適応の代表的な手法には, 解析したいデータであるターゲットデータと正解ラベルなどの多くの情報を持つソースデータとを同時にネットワークに入力してデータ間に共通する特徴を学習させる DANN などがある. 文献 [1] では合成データで動画像ドメイン適応を行っている. Kinetics-Gameplay というゲームプレイ動画から作成したデータをソースデータに利用して, ターゲットデータ Kinetics の 50 のサブクラスの分類に 17.22% から 27.50% の精度向上を達成した. しかしながら, この精度はラベルを使用してターゲットデータで学習した場合の 64.49% に対し不十分である.

我々は, 合成動画像データを活用した教師なし学習による高精度な実動画像データ解析の実現のため, 写実的な合成動画像データを作成して実験したが, 合成データのみでの学習でもドメイン適応を用いた学習でも十分な解析精度が得られず, 改善の余地があることがわかった [3]. 本研究では, より時間的モデリングに優れたドメイン適応を行うモデルである TRN, TA^{3}N を用いた手法で改善を図り, 高精度な動画像分類を実現した.

図1　Ochahouse-Syn

図2　Ochahouse-Real

2　合成動画像データセットと解析モデル

我々は, 合成動画像におけるドメイン適応のための Ochahouse Dataset を作成した [3]. これは部屋の中を 1 人の人が行動する様子を 1 台の固定されたカメラで収録した合成動画像 Ochahouse-Syn と, 実動画像 Ochahouse-Real により構成される. Ochahouse-Syn の作成には Unity® を使用した. Ochahouse-Real は, 実験住宅 OchaHouse[4] 内で筆者が動作を行い収録した. Ochahouse Dataset では, walking, sitting down, sitting, standing up, lying down, lying, getting up の 7 種類の動作クラスを作成した. 各動作クラスのデータ数は表 1の通りであり, 各動画像は約 3 秒から 7 秒程度の長さとなっている. 作成した動画像データの 1 フレームを図 1, 図 2に示す.

本研究で用いた動画像解析モデル TRN (Temporal Relation Network)[2], TA^{3}N (Temporal Attentive Adversarial Adaptation Network) について説明する. TRN は 2 次元畳み込みを行う動画像解析モデルの 1 つであり, 複数の RGB 画像フレームから関係推論を行うことで, 高精度な動画像分類を達成する. TA^{3}N (Temporal Attentive Alignment Adversarial Network)[1] は TRN を Attention 機構及び DANN によるドメイン適応で拡張したものであり, 時間的モデリングに優れた教師なし動画像ドメイン適応モデルである.

3　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

A Examination of Utilization of Synthetic Video Data for Action Recognition using Domain Adaptation
†Hana Isoi
‡Atsuko Takehusa
§Hidemoto Nakada
†Masato Oguchi
†Ochanomizu University
‡National Institute of Informatics
§National Institute of Advanced Industrial Science and Technology (AIST)

表1　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 Ochahouse-Syn	997	747	1118	780	250	250	250
実データ Ochahouse-Real	96	44	56	51	32	39	32

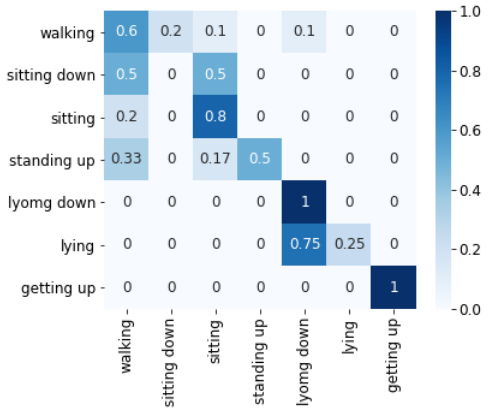

図3　TA^{3}N での混同行列

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 を使用した．各モデルはすべてベースを ResNet-18 とした．計算には 1 台の Tesla V100 PCIe 32GB を用いた．

各モデルで, 実データのみを用いた学習 (教師あり学習), 合成データのみを用いた学習 (教師なし学習), 両データを用いた実データの正解ラベルは使わないドメイン適応を行う学習 (教師なしドメイン適応) による動作分類精度を表 2 に示す．表 2 より, 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教師あり学習と同程度の精度で動作分類できており, これらに含まれる時間関係推論を行うモジュールがデータ間のドメインシフトの解消に有効であること, Attention 機構は有効でないことがわかった．

また, TA^{3}N での分類結果の混同行列を図 3に示す．図 3から, 動作クラス sitting down 及び standing up を walking , lying を lying down へという, 似た動作の特徴を持つ動画像の誤分類が多いことがわかる．また, 動作クラス walking への誤分類が多いことから, 学習データの分布を反映している可能性がある．

表2　実データ動作分類精度 (%)

学習方法	教師あり	教師なし	教師なしドメイン適応
学習データ	実データ	合成データ	実データ + 合成データ
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

4　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, 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が有効であることがわかった．

今後はより効果的に DANN と TRN を組み合わせた新た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ー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．

謝辞

この成果の一部は，JSPS 科研費 JP19H04089, JP19K11994 及び，2021 年度国立情報学研究所公募型共同研究 (21S0602) の助成を受けたものです．

参考文献

[1] Chen, M.-H., Kira, Z., AlRegib, G., Yoo, J.,Chen, R. and Zheng, J.: Temporal Attentive Alignment for Large-Scale Video Domain Adaptation, ICCV2019
[2] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reason-ing in videos, ECCV 2018
[3] 礒井葉那, 竹房あつ子, 中田秀基, 小口正人: 動作認識のための合成データ活用に向けたドメイン適応手法の比較, DICOMO2021
[4] OchaHouse Project Page, http://is.ocha.ac.jp/~siio/index.php?OchaHouse