DEIM Forum 2022 D34-1

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　〒 112–8610 東京都文京区大塚 2–1–1
†† 国立情報学研究所　〒 101–8430 東京都千代田区一ツ橋 2–1–2
††† 産業技術総合研究所　〒 305–8560 茨城県つくば市梅園 1–1–1
E-mail: †{hana,oguchi}@ogl.is.ocha.ac.jp, ††takefusa@nii.ac.jp, †††hide-nakada@aist.go.jp

あらまし　ディープニューラルネットワークの利用に伴う実データ収集のコストやプライバシーの問題に対応するため, 人工的に作成される合成データを学習に活用することが期待される. 我々は, 実データ動作認識のための写実的な合成動画像データを作成し, ドメイン適応ネットワーク DANN を 2 つの基本的な動画像識別ネットワーク 3D ResNet, TSN で拡張した動画像ドメイン適応手法の効果を調査したが, 十分な精度で実データの動作分類を行うことができなかった. 本研究では, DANN モデルのバックボーンの比較, Attention の導入, 敵対的学習の追加を行い, さらに詳細に動画像ドメイン適応手法を検討する. 実験から, 合成データをソースデータとするドメイン適応によって高精度なラベルなし実データの動作分類ができるようになることを示す.

キーワード　ドメイン適応, 合成データ, 動作分類

1　はじめに

ディープニューラルネットワーク (DNN) 技術の発展により, コンピュータービジョン分野でも DNN を用いた様々な研究がなされている. 動画像から人間の行動を DNN で解析する研究も複数あり [1], [2], [3], 家庭内の子供や高齢者の見守りなどへの応用が期待されている. DNN による画像解析の学習精度は, ラベル付き学習データセットのサイズとバリエーションに大きく依存していることが知られている [4] が, 十分なデータの収集とラベル付けには大変な時間と費用がかかる. また, 家庭内などの状況では, プライバシを保持しつつ学習データを収集するのが困難である.

そうした学習データ不足の問題に対する解決策として, 合成データの活用や, 既存データセットによるドメイン適応が有効であることが知られており, 画像解析分野で様々な研究が行われている [5], [6], [7]. 合成データはコンピュータを用いて生成されるデータのことである. 合成データは実データと比較して大量かつ多様なデータを容易かつ安価に生成できるうえに, 人手より正確なラベル付けができるという利点をもつ. 実データに混ぜて, あるいは合成データのみで学習データとして利用することができる. ドメイン適応は, 学習データとテストデータの分布が異なるときに解析精度が低下してしまうドメインシフトに対処する手法である. 特に, 合成データのみで学習した DNN の実データ解析への利用時には必須となる. さらに, ラベル付きデータをソースデータに用いることでラベルなしデータを効果的に学習する教師なしドメイン適応は, 収集が困難な実世界のデータ解析を解析する場合に有用である.

しかし, 静止画像や実動画像データのドメイン適応と比較して, 合成動画像データでのドメイン適応 に関する研究はあまり行われておらず, その成果は十分でない. また研究のためのデータセットも充実していない. そのため, 動画像解析に有用な合成動画像の特徴は明らかになっておらず, またドメイン適応の成果が十分でない原因は明らかでない. Chen らは, ゲームプレイ動画から切り抜いた動作を収録した合成動画像データセット Kinetics-Gameplay を作成し, 動作分類を行った [8]. しかし, Kinetics-Gameplay に収録されている動画像は写り方やカメラ角度が限定的で実データとは異なっており, 彼らの提案するドメイン適応手法を用いて行った動作分類では十分な精度が得られていない.

我々は, 見守りサービスのように動画像学習データの収集が困難な状況でも効果的な学習が行えるようにするため, 動作分類のための合成動画像データセット Ochahouse Dataset を作成した [9].

文献 [9] で Ochahouse Dataset を用いて行った実験では, 合成データをソースデータとするドメイン適応ネットワーク DANN により正解ラベルなし実データの分類精度が向上したが, 十分に高い精度は得られなかった.

本稿では, 更に詳細にさまざまな学習方法を調査し, 十分な精度で実データの動作分類ができるような合成データによる動画像ドメイン適応手法を明らかにする. 実験ではまず, DANN によるドメイン適応に最適な特徴抽出方法を明らかにし, さらに Attention を行い, DANN による敵対的学習を行う箇所を追加する効果を確認する. また, 先行研究 [8] で提案された既存モデル TA^{3}N との違いを比較する. その結果, 時間関係推論を行う手法である TRN で時間関係特徴の抽出を行うこと, 複数箇所の敵対的学習とドメイン推定値の確信度に基づく Attention を行い効果的なドメイン適応を行うことで, 実データ・合成デー

図 1　実動画像 Ochahouse-Real の 1 フレーム

図 2　合成動画像 Ochahouse-Syn の 1 フレーム

タ間のドメインシフトを解消でき, 実データの正解ラベルを用いない学習でも, 高精度な実データの分類ができることを示す.

2　関 連 研 究

2.1　動画像分類

動画像分類では, 画像分類で行われるような 2 次元畳み込みによる空間表現の取得に加えて, 時間表現を取得する必要がある. 動画像解析ネットワークは, 主に 3 次元畳み込みを行うもの [10], [11], [12], [13] と, 2 次元畳み込みを行うものに分類される. I3D [14] や 3D ResNet [13] を代表とする 3 次元畳み込みネットワークは通常の空間情報の畳み込みと同時に時間情報の畳み込みを行う手法である. また, 時間的な情報を補うためにオプティカルフローを用いる 2 次元畳み込みを行う代表的なモデルである Two-Streeam CNN [15] や, それを発展させた TSN(Temporal Segment Networks) [16] では, RGB 画像を入力として空間的な学習をする CNN とオプティカルフローから時間的な学習をする CNN を組み合わせ, 動画像の解析を行う. (2+1)D CNN [17] は, 3D CNN と 2D CNN の計算コストと精度のトレードオフをとったものである. また近年では, 2D CNN から時間情報の取得に着目して発展させた SlowFast [18], TANet (Temporal Adaptive Networks) [19], TRN (Temporal Relation Networks) [20] や, TSM (Temporal Shift Module) [21], Attention[17] を活用するものなどがある.

2.2　合成データ

合成データとはコンピュータにより人工的に生成されるデータであり, 実データと比較して大量かつ多様な生成が容易であるという利点をもつ. 合成データは, 実データに加えて学習データの多様化・増量や, ドメイン適応を含む転移学習に活用される.

合成データのみでの学習は, ロボットのシミュレーション実験などを主な目的として研究されてきた. 文献 [5] では, ImageNet での事前学習後に合成画像のみを用いてファインチューニングされたニューラルネットワークにより, 単純なロボット制御が可能であると示された. 2017 年には, Tobin らが合成データのテクスチャ, オクルージョンレベル, シーンの照明やカメラの視野, ノイズをランダムに変更するドメインランダム化を行うことで, ドメイン適応を行わずに単純な合成画像のみで学習したニューラルネットワークで実画像の高精度な物体検出に初めて成功した [6].

実データを増強する目的で合成データを用いる研究も複数ある. Virtual KITTI [7] は, 実際の都市での運転シーンにおける物体検出のための合成動画像データセットとして作成された. このデータセットはカメラの視点, 光源, オブジェクトのプロパティをランダム化した写実的な画像から構成され, 合成データが物体検出, 特にマルチオブジェクトの追跡において実世界の解析に有用であることを示した.

動画像における合成データに関する先行研究には [22], [23], [8] がある. 文献 [22] では, 多様で写実的な合成人間行動動画像のデータセット PHAV (Procedural Human Action Videos) を作成し, PHAV を人間行動実データセットである HMDB-51 [24], UCF-101 [25] に加えて学習すると解析精度が向上することを示した. 文献 [23] では, 動画像のテキストや背景は物体の動きを表現するオプティカルフローにほとんど影響を与えないことに着目し, 背景を簡略化した人間行動合成動画像データセットを作成した. このデータセットから抽出したオプティカルフローで, RGB 画像に加えてオプティカルフロー画像を学習に利用する動画像解析ネットワークの 1 つ TSN を追加で学習することにより, UDF-101 および HMDB-51 における精度向上に有効であることを示した.

合成データのみで構成される動画像データセットでは, Kinetics-Gameplay [8] がある. Kinetics-Gameplay は, ゲームプレイ動画像から収集した 50 クラスの行動データセットであり, ドメイン適応による動作分類のために作成された. Kinetics と重複する 50 クラスの学習において, 合成データのみでの学習よりもこのデータセットをドメイン適応に使用した学習を行うことで動作分類精度が向上することが示されたが, 十分な精度は得られていない.

2.3　ドメイン適応

ドメイン適応とは, 合成データと実データのように学習データとテストデータの特徴量分布が異なるドメインシフトに対応するための手法である. ドメイン適応の代表的な手法には, DANN (Domain-Adversarial Neural Networks) [26] や ADDA (Adversarial Discriminative Domain Adaptation) [27] などがある. DANN 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敵対的学習を行いデータ間に共通する特徴を学習させる手法である. ADDA は, ソースデータで学習させた特徴抽出器をターゲットデータ用特徴抽出の学習に用いる手法である.

動画像におけるドメイン適応では, Pan らはクロスドメイン共同 Attention 機構を提案し, ドメイン間の時間的なずれの問題に対処する方法を提案した [28]. また Choi らは, より識別性の高いクリップに焦点を当て, ビデオレベルのアラインメントを直接最適化する Attention メカニズムを提案した [29]. さらに, 補助タスクとしてクリップ順序予測を使用し, これらにより行動に大きく関与している人物や物体に焦点を当てた表現を学習することに成功した.

また, Chen らは合成データセット Kinetics-Gameplay の提案と同時に TA^{3}N (Temporal Attentive Adversarial Adaptation Network) というドメイン適応ネットワークを提案してい

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

る [8]．TA^{3}N では, 学習と同時に動き情報のアラインメントを行い, さらに Attention 機構を用いて動きの情報を明示的に考慮することで, 実データ間において高精度なドメイン適応を実現した．このネットワークでは合成動画像データセット Kinetics-Gameplay をソースデータに利用した合成動画像のドメイン適応により, ターゲットデータである Kinetics の 30 のサブクラスの分類に 17.22% から 27.50% まで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精度 64.49%に対して不十分であり, 合成動画像のドメイン適応による高精度な実データ解析は課題として残されている．

3　Ochahouse Dataset

我々は, 合成動画像におけるドメイン適応手法の有効性を調査するために利用可能な Ochahouse Dataset を作成した [9]．このデータセットは, 部屋の中を 1 人の人が自由に動きまわり 7 種類の動作をする様子を 1 台の固定されたカメラで収録した実動画像 Ochahouse-Real と合成動画像 Ochahouse-Syn で構成される．作成した動画像データの 1 フレームを図 1, 図 2 に示す．Ochahouse-Real は, お茶の水女子大学の実験住宅 OchaHouse [30] 内で筆者が動作を行い収録した．Ochahouse-Syn はゲームエンジン Unity® を使用して作成し, OchaHouse を模擬した室内で人間モデルが動作する様子を収録した．Unity® は無料でゲームや動画像を作成することができるオープンなプラットフォームであり, 多くのユーザによるコミュニティの活動が盛んである．文献 [22] と同様に, 我々は人間のアセットとモーションキャプチャから作成された行動アニメーションを Unity Asset Store から入手した．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 秒程度の長さである．また, いずれもフレームレートは 5fps とした．

作成した Ochahouse-Syn は Ochahosue-Real と見た目が非常に類似しているにもかかわらず, 我々が文献 [9] で行ったドメイン適応の効果の実験では, Ochahouse-Syn を活用する Ochahouse-Real の正解ラベルを使わない動画像ドメイン適応では, 十分な精度での Ochahouse-Real の動作分類は達成されなかった．

4　ドメイン適応手法の比較実験

我々は文献 [9] では合成データをソースデータとするドメイン適応を行う動画像学習による高精度な実データの分類が実現できなかったため, より詳細に効果的な動画像学習手法を調査する．文献 [9] では 3D ResNet, TSN, の線形層の前までをそれぞれ特徴抽出器とした DANN で精度を調査したが, 本研究ではさらに ResNet, TRN, TANet の線形層の前までをそれぞれ特徴抽出器として用いた DANN の精度を比較し, 効果的なモデルのバックボーンを明らかにする．次に, モデルに Attention 機構を導入し, その効果を確認する．最後に, DANN の DA Block の場所・個数を変化させ, 最適なモデルの構造を調査する．

4.1　バックボーンモデルの選定

DANN によるドメイン適応を行うモデルの特徴抽出器を動画像分類モデル 3D ResNet, ResNet, TSN, TRN, TANet のそれぞれの線形層以前の部分として学習し, 実データの分類精度を比較する．

DANN は最も基本的なドメイン適応手法である．図 3 にバックボーンを 3D ResNet とするドメイン適応モデル（3D ResNet+DANN）を示す．DANN では, 各バックボーンモデルから得られた動画像の特徴量から, クラス分類器 F と DA Block (Domain Adversarial Block, ドメイン分類器, または弁別器) で, それぞれクラス分類ロスとドメイン分類ロスが出力され, それらの荷重和を最小化する学習を行う．すなわち, 以下の最適化を行う．

$$\min \mathcal{L}_{class} + \alpha \mathcal{L}_{domain} \tag{1}$$

$$= \min L_c(\hat{\boldsymbol{y}}, \boldsymbol{y}) + \alpha L_d(\hat{\boldsymbol{d}}, \boldsymbol{d}) \tag{2}$$

$$= \min L_c(F(E(\boldsymbol{x})), \boldsymbol{y}) + \alpha L_d(D(GRL(E(\boldsymbol{x}))), \boldsymbol{d}) \tag{3}$$

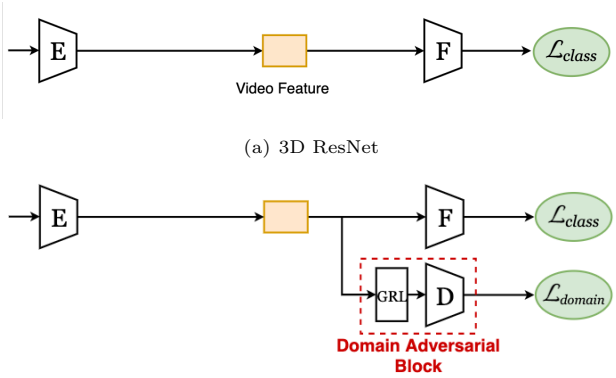
$\boldsymbol{y}, \boldsymbol{d}$ はそれぞれ, クラスラベルとドメインラベルである．α はドメイン分類器の学習をスケジューリングするスカラー値であり, 学習の序盤は 0 で, 学習が進むにつれ 1 に近づく．また GRL (Gradient Reverse Layer) では以下のような計算が行われる．

$$GRL(\boldsymbol{x}) = \boldsymbol{x} \tag{4}$$

$$\frac{dGRL}{d\boldsymbol{x}} = -\rho \boldsymbol{I} \tag{5}$$

ドメイン分類損失 $\mathcal{L}_{domain}$ を最小化する学習を進めると, よりドメインの情報を保持しない内部表現を得ることができるようになる．すなわち, ソースデータかターゲットデータかに依存しない動作の情報を得るようになる．

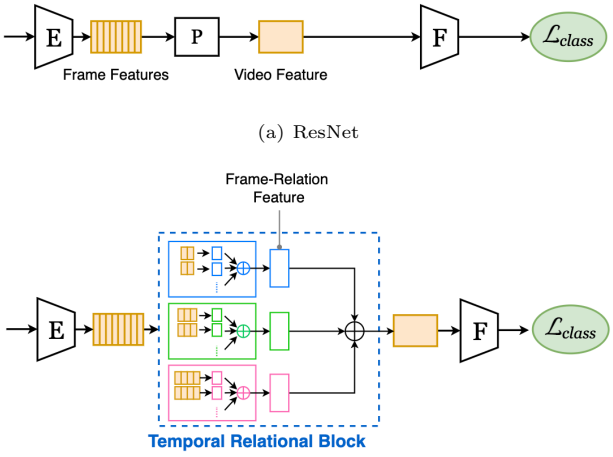
3D ResNet バックボーンでは, 動画像から切り出した複数枚の RGB フレームに 3 次元畳み込み処理を行い, 動画像の特徴を得る．ResNet バックボーンでは, 各画像フレームからそれぞれフレームごとの特徴量を抽出し, それらを pooling 層で集約する．TSN バックボーンは ResNet の拡張であり, 動画像の RGB フレームを解析する spatial stream, flow フレームを解析する temporal stream の 2 つの畳み込みネットワークから構成される．TSN ではオプティカルフローを用いるので, 他の手法との比較のために TSN の RGB フレームを解析する spatial stream を TSN-spatial, オプティカルフローを解析する



(a) 3D ResNet

(b) 3D ResNet + DANN

図 3　(a)3D ResNet と, (b)3D ResNet をバックボーンとする DANN. E は 3 次元畳み込みを行う畳み込み層からなる特徴抽出器, F,D はそれぞれ線形層からなるクラス分類器とドメイン分類器, GRL は勾配反転層.



(a) ResNet

(b) TRN

図 4　(a) ResNet と (b) TRN. ともに E は ResNet-18 の畳み込み層で構成される特徴抽出器. P は Pooling 層. TRN では Temporal Relational Block により時間関係推論を行う.

temporal stream を TSN-temporal として実験に用いた. TRN バックボーンでは ResNet でフレームごとの特徴量を抽出した後に Temporal Relation Module で関係推論を行い, フレーム関係特徴量を得る. TANet バックボーンでは Attention 機構を組み込んだ ResNet により, 動画像の特徴量を得る.

実験に用いるすべてのバックボーンモデルは ResNet-18 をベースに構成し, F には ReLU, Dropout を含む 3 層の線形モデルを, D は ReLU を含む 2 層の線形モデルを, 損失関数 L_c, L_d にはクロスエントロピー誤差関数を用いた.

実験では, それぞれ異なるバックボーンをもつ複数のモデルで Ochahouse Dataset を用いた動作分類を行い, 結果を比較する. 各モデルで, 正解ラベル付き実データ Ochahouse-Real のみでの学習, 正解ラベル付き Ochahouse-Syn のみでの学習, ソースデータとしてラベル付き合成データ Ochahouse-Syn を, ターゲットデータとしてラベルなしの実データ Ochahouse-Real を用いる教師なしドメイン適応を行う学習を行い, Ochahouse-Real の解析精度を調査する. 3D ResNet ではすべての動画像フレームを, TSN では 3 枚, TRN では 5 枚のフレームを入力に用いた. 最適化手法には SGD（勾配降下法）を, バッチサイズはドメイン適応しない場合は 16, ドメイン適応時は各ドメインで 16 ずつ合計 32 として 60 エポック学習した. 計算には 1 台の Tesla V100 PCIe 32GB を搭載した計算機を用いた.

表 2　各バックボーンでの実データ動作分類精度 (%)

バックボーン	実データ	合成データ	両データでドメイン適応
3D ResNet	88.60	11.11	40.74
ResNet	65.56	28.38	46.75
TSN	78.13	25.64	37.27
TSN-spatial	78.13	25.64	37.27
TSN-temporal	79.69	26.92	34.71
TRN	66.67	46.41	63.08
TANet	94.02	20.23	27.21

各学習手法で学習した時の実データ Ochahouse-Real の動作分類精度を表 2 に示す. 表 2 の「実データ」の列は実データで教師あり学習をしたそれぞれのモデルでの実データの分類精度を,「合成データ」の列は合成データで教師あり学習をしたそれぞれのモデルでの実データの分類精度を,「両データでドメイン適応」の列はラベルなし実データをターゲットデータ, ラベル付き合成データをソースデータとするドメイン適応を行うそれぞれのモデルをバックボーンとする DANN での実データの分類精度を表す. 表 2 から, すべてのモデルにおいて実データでの学習に対し, 実データの正解ラベルを使わない学習を行う手法である「合成データ」「両データでドメイン適応」では分類精度が大幅に低く, ドメインシフトが起こっていることがわかる. また,「合成データ」での学習よりも,「両データでドメイン適応」を行う学習のほうが精度が高く, ドメインシフトの解消に DANN によるドメイン適応が有効であることがわかる. また, バックボーンとして TRN を用いたモデルの分類精度が最も高いことから, バックボーンには TRN が最適であることがわかった. さらに, TSN, TSN-spatial, TSN-temporal をバックボーンとするモデルでの分類精度の比較から, TSN-temporal の精度が低くオプティカルフロー入力が有効でないことがわかった.

4.2　Attention 機構の導入

先行研究 [8] に基づき, Temporal Relational Block で計算されたフレーム関係の特徴量に重みをつける Attention 機構をモデルに導入し, その効果を確認する. Attention 機構は以下でその重みが定義される General Attention と, 先行研究 [8] で提案された Domain Attention の 2 種類を用いた.

General Attention では, 重み w は以下で計算される.

$$\boldsymbol{w} = softmax(\phi(tanh(\psi(\boldsymbol{r})))) \tag{6}$$

ここで, $\boldsymbol{r}$ はフレーム関係特徴量であり, ϕ と ψ は線形変換を行う関数である. Domain Attention では, 重み w は以下で算出される.

$$\boldsymbol{w} = 1 - H(\hat{\boldsymbol{d}}) \tag{7}$$

ここで, $H(p) = -\sum_k p_k \cdot \log(p_k)$ はエントロピーであり, ド

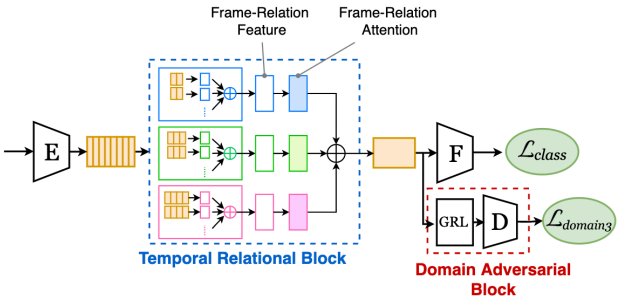


図 5　TRN バックボーンの DANN に Attention 機構を導入したモデル

表 3　TRN バックボーンの DANN に Attention 機構 を導入したモデルでの実データ動作分類精度 (%)

Attention 機構	実データ	合成データ	両データでドメイン適応
なし	66.67	38.46	52.82
General Attention	66.15	29.49	54.40
Domain Attention	66.92	34.10	60.86

メインの予測値 $\hat{d}_i^n$ の確信度が高いほど重み w_i^n は大きくなる．

TRN をバックボーンとする DANN モデルに Attention 機構を追加したモデルを図 5 に，また，その実データ動作分類精度を表 3 に示す．表 3 から，TRN ベースの DANN では，Attention 機構を加えた方が高精度になること，さらに，Domain Attention が最も高精度になることがわかった．

4.3　DA Block の個数の変化と既存モデル TA^{3}N との比較

先行研究 [8] に基づき，さらに 2 つの Domain Adversarial Block を導入し，効果を調査する．Domain Adversarial Block を 3 つ導入したモデルを図 6 に示す．

また，Domain Adversarial Block を 3 つ導入し，さらに Attentive Entropy Loss の最小化も行う既存モデル TA^{3}N [8] との比較を行う．Attentive Entropy Loss は以下で定義される損失であり，この最小化によってドメインの予測値とクラスの予測値のエントロピーを最小化する．

$$L_{ae} = (1 + H(\hat{\boldsymbol{d}}_3)) \cdot H(\hat{\boldsymbol{y}}) \tag{8}$$

$\boldsymbol{d}_3$ は Domain Adversarial Block 3 で計算されるドメイン予測値である．

追加したそれぞれの Domain Adversarial Block と既存モデル TA^{3}N の効果を確認するために，図 6 の Domain Adversarial Block 1 と Domain Adversarial Block 2 の適用の有無を変化させたモデルと TA^{3}N での実データ分類精度を表 4 に示す．表 4 では，”o” または ”-”はモデルの Domain Adversarial Block (DA Block) 1, 2, 3 のそれぞれの適用の有無を表しており，”TA^{3}N [8]”は既存モデル TA^{3}N を表す．表 4 から，適用する Domain Adversarial Block の個数が多いほうが高精度になることがわかる．また，Attentive Entropy Loss の最小化を追加する拡張をした TA^{3}N ではわずかに精度が低下しており，学習時のドメイン予測値とクラス予測値のエントロピーの最小化は効果が確認できなかった．

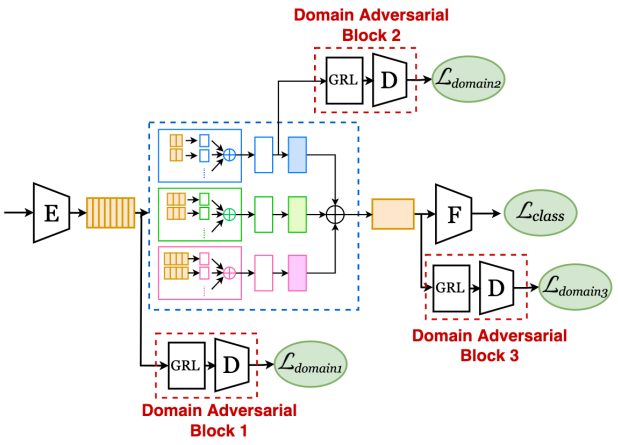


図 6　TRN バックボーンの DANN に ATtention, 複数箇所の Domain Adversarial Block を導入したモデル

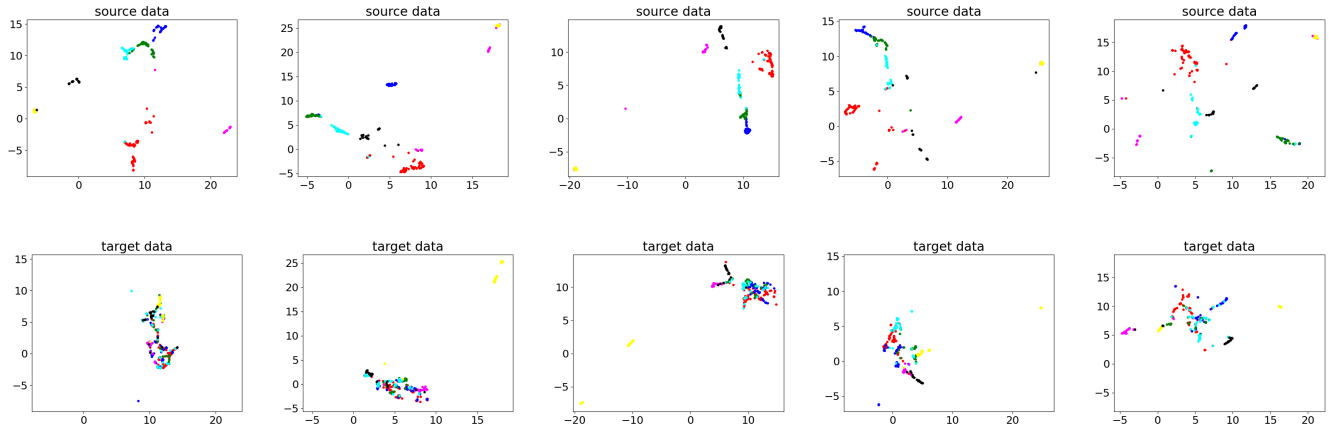
表 4　複数の Domain Adversarial Block を導入したモデルでの実データ動作分類精度 (%)

DA BLock 1	DA BLock 2	DA BLock 3	両データでドメイン適応
-	-	o	60.43
o	-	o	55.28
-	o	o	57.88
o	o	o	61.43
TA^{3}N [8]			58.42

4.4　考　　察

それぞれのモデルの特徴抽出器により抽出される動画像特徴量を 2 次元に圧縮しプロットすることで，4.2 節と 4.3 節で行ったモデルの拡張の効果を考察する．図 7 に，TRN，TRN + DANN，TRN + DANN + Domain Attention，TRN + 3 箇所の DANN + Domain Attention，TA^{3}N で抽出される Ochahouse-Syn，Ochahouse-Real それぞれの動画像特徴表現ベクトル V を UMAP [31] で 2 次元に圧縮しプロットしたものを示す．図 7 の各色は赤が walking，青が sitting down，緑が sitting，シアンが standing up，マゼンタが lying down，黄が lying，黒が walking の動作クラスを表している．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．各点が色ごとに分かれて固まって配置されている場合は，クラス分類ができるような特徴表現ベクトルの抽出に成功していることを意味する．Source data と target data で点が同様な位置に配置されている場合は，これらのデータからドメインによらない特徴表現ベクトルが抽出できていることを意味する．ドメインシフトを解消するには，ドメインによらない，かつクラス分類のための動作の特徴を保持するような特徴表現ベクトルを抽出すること，すなわち source data と target data の各点の分布が色ごとに分かれて固まって，同様な位置に配置されている必要がある．

図 7(a) では，ドメイン適応を行わない TRN で抽出された動画像特徴表現ベクトルは source data と target data で分布の形状が大きく異なっていることからドメインシフトが起こって



(a) TRN　(b) TRN+DANN　(c) (b) + Attention　(d) (c) + 複数箇所の DANN　(e) TA^{3}N

図 7　(a) TRN, (b) TRN + DANN, (c) TRN + Domain Attention + DANN, (d) TRN + Domain Attention + 3 箇所のの DANN, (e) TA^{3}N の各モデルで抽出される 合成動画像と実動画像の特徴ベクトル $\boldsymbol{V}$ の UMAP による可視化.

いることがわかる. また, 図 (a) から (b), (c), (d) とモデルを拡張するにつれ, source data と target data の特徴量の分布がより近づいていることから, モデルを拡張することでより効果的にドメインシフトの解消ができるようになっていることがわかる. ただし, いずれのモデルでも source data と target data の分布は一致しておらず, このことはドメイン適応手法に改善の余地があることを示唆する.

これらの実験結果から, 写実的に作成された合成データをソースデータとして用いるドメイン適応では, 1. TRN をバックボーンモデルとすること, 2. ドメイン推定値の確信度に基づく Attention 機構を追加すること, 3. Domain Adversarial Block でのドメイン適応は複数箇所で行うこと, が有効であることが示された. また, 先行研究で行われているドメイン推定値・クラス推定値のエントロピーの最小化は効果が確認できなかった. 1. は, Ochahouse Dataset の合成データと実データの違いは動きの速さやタイミングにあるため, TRN による時間関係推論で動画像のフレームごとのタイミングの差を解消することで, ドメインシフトを大幅に減らすことができることを示唆する. また, Domain Adversarial Block の学習を効果的に行うためには 2. と 3. の方法が有効であるが, さらに改善できる可能性があることがわかった. これらの結果から, これらの手法によって, 合成データをソースデータとして活用したドメイン適応を行うことにより, 正解ラベルがない実データの動作分類ができるようになることが明らかになった.

5　まとめと今後の課題

我々は, 作成した Ochahouse Dataset を用いた実験により, 写実的な合成データをソースデータとして動画像ドメイン適応を行う学習により, 正解ラベルのない実動画像データの高精度な動作分類ができるようになることを示した. 我々が作成した屋内での見守り・監視を想定したドメイン適応のためのデータセット Ochahouse Dataset は, 人間の動きを記録した実動画像 Ochahouse-Real と, それを写実的に模した Ochahouse-Syn が含まれている. 本研究では, これらを用いた実験で, 合成動画像データと実動画像データの間のドメインシフトの問題は, TRN バックボーンでの時間関係推論によりソースデータとターゲットデータのフレーム間の差を解消すること, DANN とドメイン推定値の確信度に基づく Attention を導入し効果的なドメイン適応を行うことで解消でき, 実データの正解ラベルを用いない学習でも, 高精度な実データの分類ができるようになることを示した. これらの結果から, プライベートな空間で記録されたシーンのようなラベル付き実働画像データの収集が困難な場合でも, 動画像データを使用した高齢者や子供の自動モニタリングシステムの実現に応用できる可能性があることがわかった.

今後は, より高精度な解析の実現を目指し, また, 物体検出や異常検知などのより多くのビジョンタスクへの応用を目指す.

謝　　辞

この成果の一部は, JSPS 科研費 JP19H04089, JP19K11994, JP18K11488 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.

文　　献

[1] Cheng, G., Wan, Y., Saudagar, A. N., Namuduri, K. and Buckles, B. P.: Advances in Human Action Recognition: A Survey, *ArXiv*, Vol. abs/1501.05964 (2015).

[2] 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).

[3] Takasaki, C., Takefusa, A., Nakada, H. and Oguchi, M.: A Study of Action Recognition Using Pose Data Toward Distributed Processing Over Edge and Cloud, *2019 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, pp. 111–118 (online), 10.1109/CloudCom.2019.00027 (2019).

[4] Sun, C., Shrivastava, A., Singh, S. and Gupta, A.: Revisit-

ing unreasonable effectiveness of data in deep learning era, *Proceedings of the IEEE international conference on computer vision*, pp. 843–852 (2017).

[5] Sadeghi, F. and Levine, S.: CAD^{2}RL: Real Single-Image Flight without a Single Real Image, *ArXiv*, Vol. abs/1611.04201 (2016).

[6] Tobin, J., Fong, R. H., Ray, A., Schneider, J., Zaremba, W. and Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world, *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23–30 (2017).

[7] Gaidon, A., Wang, Q., Cabon, Y. and Vig, E.: Virtual Worlds as Proxy for Multi-object Tracking Analysis, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4340–4349 (2016).

[8] Chen, M.-H., Kira, Z., AlRegib, G., Yoo, J., Chen, R. and Zheng, J.: Temporal Attentive Alignment for Large-Scale Video Domain Adaptation, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2019).

[9] 儀井葉那, 竹房あつ子, 中田秀基 and 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, **マルチメディア，分散，協調とモバイル（*DICOMO 2021*）シンポジウム** (2021).

[10] Ji, S., Xu, W., Yang, M. and Yu, K.: 3D Convolutional Neural Networks for Human Action Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 1, pp. 221–231 (online), 10.1109/TPAMI.2012.59 (2013).

[11] Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M.: Learning Spatiotemporal Features with 3D Convolutional Networks, *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 4489–4497 (online), 10.1109/ICCV.2015.510 (2015).

[12] Carreira, J. and Zisserman, A.: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4733 (2017).

[13] Hara, K., Kataoka, H. and Satoh, Y.: Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition, *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pp. 3154–3160 (2017).

[14] Carreira, J. and Zisserman, A.: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (online), 10.1109/cvpr.2017.502 (2017).

[15] Simonyan, K. and Zisserman, A.: Two-Stream Convolutional Networks for Action Recognition in Videos, *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, Cambridge, MA, USA, MIT Press, p. 568–576 (2014).

[16] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X. and Van Gool, L.: Temporal Segment Networks: Towards Good Practices for Deep Action Recognition, *Computer Vision – ECCV 2016* (Leibe, B., Matas, J., Sebe, N. and Welling, M., eds.), Cham, Springer International Publishing, pp. 20–36 (2016).

[17] Tran, D., xiu Wang, H., Torresani, L., Ray, J., LeCun, Y. and Paluri, M.: A Closer Look at Spatiotemporal Convolutions for Action Recognition, *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6450–6459 (2017).

[18] Feichtenhofer, C., Fan, H., Malik, J. and He, K.: SlowFast Networks for Video Recognition, *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6201–6210 (online), 10.1109/ICCV.2019.00630 (2019).

[19] Liu, Z., Wang, L., Wu, W., Qian, C. and Lu, T.: TAM: Temporal Adaptive Module for Video Recognition, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13708–13718 (2021).

[20] Zhou, B., Andonian, A., Oliva, A. and Torralba, A.: Temporal Relational Reasoning in Videos, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I* (Ferrari, V., Hebert, M., Sminchisescu, C. and Weiss, Y., eds.), Lecture Notes in Computer Science, Vol. 11205, Springer, pp. 831–846 (online), 10.1007/978-3-030-01246-5_49 (2018).

[21] Lin, J., Gan, C. and Han, S.: TSM: Temporal Shift Module for Efficient Video Understanding, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2019).

[22] De Souza, C. R., Gaidon, A., Cabon, Y. and López, A. M.: Procedural Generation of Videos to Train Deep Action Recognition Networks, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2594–2604 (online), 10.1109/CVPR.2017.278 (2017).

[23] Ballout, M., Tuqan, M., Asmar, D., Shammas, E. and Sakr, G.: The benefits of synthetic data for action categorization, *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8 (online), 10.1109/IJCNN48605.2020.9207337 (2020).

[24] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T. and Serre, T.: HMDB51: A Large Video Database for Human Motion Recognition, pp. 2556–2563 (online), 10.1109/ICCV.2011.6126543 (2011).

[25] Soomro, K., Zamir, A. R. and Shah, M.: UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild, *CoRR*, Vol. abs/1212.0402 (online), http://dblp.uni-trier.de/db/journals/corr/corr1212.htmlabs-1212-0402 (2012).

[26] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M. and Lempitsky, V.: Domain-Adversarial Training of Neural Networks, *J. Mach. Learn. Res.*, Vol. 17, No. 1, p. 2096–2030 (2016).

[27] Tzeng, E., Hoffman, J., Saenko, K. and Darrell, T.: Adversarial Discriminative Domain Adaptation, *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2962–2971 (online), 10.1109/CVPR.2017.316 (2017).

[28] : Adversarial Cross-Domain Action Recognition with Co-Attention, Vol. 34, pp. 11815–11822 (online), 10.1609/aaai.v34i07.6854 (2020).

[29] Choi, J., Sharma, G., Schulter, S. and Huang, J.-B.: Shuffle and Attend: Video Domain Adaptation, *Computer Vision – ECCV 2020* (Vedaldi, A., Bischof, H., Brox, T. and Frahm, J.-M., eds.), Cham, Springer International Publishing, pp. 678–695 (2020).

[30] : OchaHouse Project Page, `http://is.ocha.ac.jp/~siio/index.php?OchaHouse`.

[31] McInnes, L., Healy, J., Saul, N. and Grossberger, L.: UMAP: Uniform Manifold Approximation and Projection, *The Journal of Open Source Software*, Vol. 3, No. 29, p. 861 (2018).