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

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@ogl.is.ocha.ac.jp, ††takefusa@nii.ac.jp, †††hide-nakada@aist.go.jp, ††††oguchi@is.ocha.ac.jp

あらまし　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に応用することが期待されている. しかし, 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ておらず, またそのようなデータセットを現実の動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試作・評価した. 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 つの動作がランダムに行われるようにした. 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. 実験より, 作成したデータにおいては上記の動作分類ができるが, 今の段階では作成したデータのみで学習した DNN では現実の動画像 STAIR-actions の動作識別はできないこと, しかし, ファインチューニングによりある程度は識別ができるようになることがわかった. また, 照明条件の変化・ノイズ・ぼかしについては解析制度への影響はほとんどみられないことがわかった.

キーワード　合成データ, 機械学習, 深層学習, 動画像, コンピュータビジョン

Development of Synthetic Video Dataset for Daily Action Classification in Living Space

Hana ISOI†, Atsuko TAKEFUSA††, Hidemoto NAKADA†††, and Masato OGUCHI†

† Ochanomizu University　2-1-1 Otsuka,Bunkyo-Ku, Tokyo 112–8610, Japan
†† National Institute of Informatics　2-1-1 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan
††† National Institute of Advanced Industrial Science and Technology (AIST)　1-1-1 Umezono, Tsukuba, Ibaraki 305–8560, Japan
E-mail: †hana@ogl.is.ocha.ac.jp, ††takefusa@nii.ac.jp, †††hide-nakada@aist.go.jp, ††††oguchi@is.ocha.ac.jp

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, 様々なコンピュータビジョンタスクへの応用に注目が集まっている. 特に, 最近の研究では動画像から人間の動作の識別や異常検知を行うことで人間の行動解析が可能になり [1] [2] [3], 家庭内の子供や高齢者の見守りなどへの活用が期待されている.

[4] が示すように, DNN によるデータ解析において, そのパフォーマンスはラベル付き学習用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. しかしながら, データの収集とラベル付けは膨大な時間と費用を要する上に, 実際に日常動作解析で必要とされるすべてのシチュエーションを収めたデータセットを作成することは非常に困難である. さらに, 個人を特定することができるようなデータの公開・利用には倫理上の問題が生じる. よって, DNN を用いたデータ解析において, 十分な学習用データセットの確保が課題となる.

学習用データセットを確保するため, 物体検出などの一部の静止画像解析タスクにおいては, シミュレーションを用いて合成画像を生成し, 学習データに用いる手法が注目されている. DNN による解析に用いるにあたり, 合成画像には現実の画像

が持つ多様な特徴を完全に表現しきれないという弱点があるが, レンダリング時に合成データの照明や天気等の様々な条件を多様化させる処理や, 現実のカメラでの画像生成時の劣化を模した加工を施すことにより, 合成データと現実のデータとのギャップを埋めることができるようになってきた [5] [6]. しかし, 動画像分類のための合成データを生成する方法に関しては, 筆者らが知る限りまだ知られていない. 特に, 本研究で目指すような室内での人間の行動を分析するには, 室内という環境に特有の問題点があるのかどうかや, カメラの条件や対象である人が不規則に動く点から, どのような合成データが人間の室内の行動解析に有用となるのか明らかでない.

本研究では, 人間の屋内での動作解析のための学習用合成動画像データを作成すること, またそのような合成動画像の生成方法を確立することを目指し, 人が歩く・立ち止まる・座る・座っている・立ち上がるの 5 つの動作をする合成動画像を試作および評価した. また, 合成動画像を現実に近づけるために, 照明条件の変化やノイズ・ぼかしの付与を施し, 動作識別精度の変化を調査した. 予備実験として, 合成データにおいて, それらの動作を 3D ResNet [10] によって識別可能であることを検証した. 3D ResNet とは, 複数の画像フレームを 3 次元のカーネルを用いて, 通常の CNN の 2 次元空間に加えて時間方向にも同時に畳み込みを行う 3D CNN の一種である. さらに, 現実の動画像データセットを用いた評価実験では, 試作した合成データセットのみで学習した DNN では現実の動作の識別はできないが, ファインチューニングによりある程度は識別できるようになることを示した. その際, 今の段階では, データに照明条件の変化・ノイズ・ぼかしを施しても, 現実のデータの解析精度への影響はほとんどみられないことが明らかになった.

本論文では以降, 2 章で合成データに関する先行研究について, 3 章では作成した合成データの概要について, 4,5 章では作成した合成データセットによる動作識別の実験と評価について, 6 章では本研究のまとめと今後の課題について述べる.

2. 関 連 技 術

2.1 ドメインランダム化

合成データによる学習では, 時刻, 天候などのドメインをランダム化してデータセットを多様化することで高い精度の学習モデルが構築できることができることが知られている.

Fereshteh Sadeghi らはシミュレーションで作成した画像のみを用いて画像からの学習を行うことを目指し, 前段階として ImageNet で事前学習し, ランダム化されたレンダリングピクセルでファインチューニングした DNN でロボット制御を行うことに成功した [7].

J tobin らは 2017 年, シミュレートするテクスチャ, オクルージョンレベル, シーンの照明, カメラの視野, レンダリングエンジン内の均一なノイズに対してドメインランダム化を行うことで, 単純な環境において, 合成画像のみで学習した DNN でドメイン適応を行わずに現実世界での高精度な物体検出に初めて成功した [5]. [5] に基づき, Adrien Gaidon ら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

図 1　座る様子（上）と立ち上がる様子（下）

”Virtual KITTI” を生成した [8]. 彼らはカメラの視点, 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をレンダリングによって生成し, 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.

ただし, これまでに行われてきたシミュレートされたデータによる解析の対象は静止画像内の物体である. 本研究では, これらの研究をふまえて, 動画像からの不規則に動く人間の行動の解析についても同様な処理が有効であるのか調査する.

2.2 センサでの劣化を模した加工

センサでのバイアスがディープニューラルネットワークの精度の低下をもたらすことがわかっている [6]. Alexandra Carlson らは, センサでの劣化を模した加工を合成データに施すことで, 現実のデータの解析精度が向上することを実験により明らかにした. 本研究でも同様に, センサでの劣化を模して, ノイズ・ぼかしを施す.

3. 作成した合成動画像

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, およびセンサの劣化を模したノイズ・ぼかし処理を行った.

3.1 Unity

行動解析のためのデータセットの作成に Unity [9] を用いる.

Unity とは, Unity Technology 社が提供するゲームエンジンであり, 無料で様々なプラットフォーム上で動かすことができるという特徴がある. Unity には独自の大きなコミュニティがあり, Asset Store でユーザが提供する多くのアセットが公開されているため, それを素材として新たな動画像の作成が容易であるという利点がある.

3.2 動　　作

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, 部屋の四隅上から 256 * 256 ピクセル, 5fps で撮影したものを用意した. 座って立ち上がる様子を図 1 に, 四隅から撮影した動画像の 1 フレームを図 2 に示す.

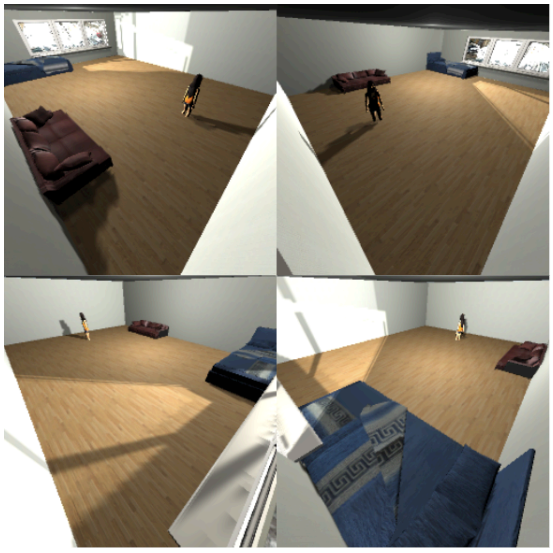

図 2　作成した動画像の 1 フレーム

3.3　照明条件のランダム化

空間内の 2 つの照明がランダムに明るさを変えながら, ランダムに移動することで, 照明条件の変更を表現した. 照明の明るさを, 部屋全体が明るく見える程度から, 薄暗く見えにくい状態まで変化させ, 照明自体も部屋の内部をランダムに移動させた.

3.4　ノ　イ　ズ

ノイズはガウスノイズを適用して式 (1) のようにモデル化した.

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

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.

3.5　ぼ　か　し

ぼかし処理はガウスフィルタを適用し, 式 (2) で表現した. $I_{blur}(x,y)$ 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.

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

ノイズ・ぼかし処理を施した 1 フレームを図 3 に示す.

3.6　部屋の内装

簡易的な内装 (a) と, より写実的な内装 (b) の 2 通りを用意した. それぞれの部屋の様子を図 4 に示す. 内装 (a) では壁や床・家具に無機質なテクスチャを使用しているのに対し, 内装 (b) では壁紙・フローリング・布を模したテクスチャを使用し, また図 4 の右側の壁に窓をつけた.

4.　予 備 実 験

実験にて, 作成した合成データを用いて動作識別ができることを確認する. また, 内装 (a) のデータを用いてデータ数, カメラの個数と精度の関係を調べる. 次に, より写実的な内装 (b) のデータを用いて, ノイズ, ぼかし, 照明条件の変更の加工を組み合わせた場合の精度を調べる.

図 3　ノイズ・ぼかしを付与した 1 フレームの例

図 4　内装 (a)(左) と内装 (b)(右)

表 1　データ数とカメラ個数を変化させた動作識別精度の比較

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%

4.1　実　験　1

作成した内装 (a) の合成データを用いて動作識別ができることを確認する. 設置するカメラの個数は各壁の中心上部 4 個, さらに四隅 4 個を追加した 8 個 の 2 通りとし, それらの学習の精度を比較する. また, 照明条件のランダム化とノイズ・ぼかし処理を施し, 動作判別への影響を調査する.

入力データは, 作成した動画像 3.2 秒分を 16 フレームに等間隔で分割したもので, それらを 10:3:3 に分割して学習データ・検証データ・テストデータとした. 各フレームに動作ラベルがついており, 最もつけられた数が多いラベルをその動画像が表す動作としている. また, ノイズ・ぼかしを施したデータでの学習時には, データに施されたノイズ・ぼかしはエポックごとにランダムな強さでかけられている. 計算には GeForce GTX 980 GPU を備えたサーバ機を用いて, 3D ResNet [10] で歩く・立ち止まる・座る・座っている・立ち上がるという 5 つの動作に分類する.

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 5 に示す. 図 5 から, データを加工せずにデータ数 4,000, カメラ個数 8 で

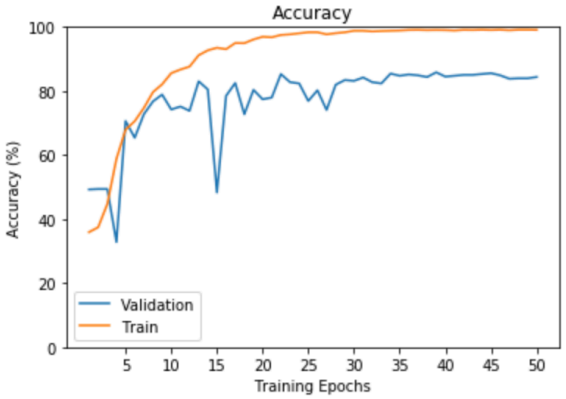

図 5　データ数 4000, カメラ数 8 での動作判別精度

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%

学習した場合は約 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を表 1 に示す．表 1 から, データ数が増加するにつれて 4 カメラ、8 カメラでも場合も動作識別精度が高くなることが分かる．データ数が同じ場合で 4 カメラ, 8 カメラの結果を比較すると、4 カメラの方が高い精度で判別できることが示された．また, カメラあたりの画像数を同じにした場合の結果, すなわちデータ数 2000, 4 カメラの結果とデータ数 4000, 8 カメラの結果の比較では同程度の精度を示していた．よって, 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示された．

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・ぼかし処理を施した結果を表 2 に示す．ノイズ・ぼかし処理を施した場合, 施さない場合と比べてほぼ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には精度が低下していることから, 照明条件を変更すると動作判別が難しくなっていることがわかるが, その場合でも約 75% の精度で動作判別ができていた．

4.2　実　験　2

作成した内装 (b) の合成動画像を用いて動作識別ができることを確認する．また, 照明条件のランダム化とノイズ・ぼかし処理による学習精度の変化を確かめるために, それぞれの有無を変更したデータで学習を行い, 動作識別への影響を調査する．

入力データは, 作成した動画像 3.2 秒分を 16 フレームに等間隔で分割したもので, 学習用に 2500 個, 検証用に 750 個用いた．各フレームに動作ラベルがついており, 最もつけられた数が多いラベルをその動画像が表す動作としている．ノイズ・ぼかしを施したデータでの学習時には, データに施されたノイズ・ぼかしはエポックごとにランダムな強さでかけられている．学習に用いたデータの照明条件の変化・ノイズ・ぼかしの有無にかかわらず, テスト時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 750 個用いた．これは, 現実の動画像データを想定したものである．これらの動画像を, 3D ResNet 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 つの動作に分類する．計算には GeForce GTX 980 GPU を備えたサーバ機を用いた．

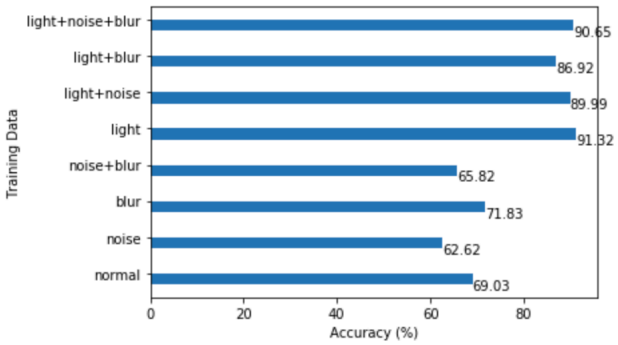

図 6　作成した合成データにおける動作識別の精度

学習データの条件を変化させて学習させ, テストを行なった結果を図 6 に示す．図の横軸は解析精度 (%), 縦軸は学習に用いた合成動画像データの種類を表す．例えば, 一番上の「light+noise+blur」は, 照明条件の変化とノイズとぼかしを全て施したデータを, 上から二番目の「light+blur」は照明条件の変化とぼかしを施したデータを, 一番下の「normal」はいずれも施していないデータを表す．図 6 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．また, 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%の精度で動作識別ができることがわかった．これらの結果から, 作成した合成データにおいて, 動作の学習ができていること, 照明条件を変化させると識別が難しくなることがいえる．

5.　STAIR-actions による評価

5.1　STAIR-actions

実写データを用いた評価のために, 本実験では STAIR-actions ビデオデータセット [11] を用いた．STAIR-actins とは, YouTube またはクラウドソースワーカから集めた 100 種類の人間の日常動作のラベル付き動画像データセットである．後述の実験には, STAIR-actions から walking, sitting down, standing up の 3 クラスを使用した．使用した動画像の 1 フレームの例を図 7 に示す．

5.2　実　験　1

本実験では, 作成した内装 (b) の合成データで学習させた 3D ResNet を STAIR-actions の動画像データ 1783 個を用いてテストし評価する．また, 照明条件のランダム化とノイズ・ぼかし処理による学習パフォーマンスの変化を確かめるために, それぞれの有無を変更したデータで学習を行い, 実写動画像の動作識別への影響を調査する．合成データでの学習の条件は, 予備実験と同様である．

実験結果を図 8 に示す．図 8 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識別するのと同等な 20%程度の精

図 7　STAIR-actions ビデオデータの 1 フレームの例（上から *walking*, *sittingdown*, *standingup*）

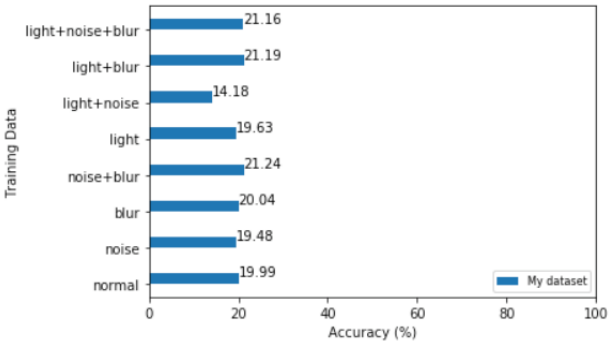

図 8　STAIR-actions による動作識別の精度

度となり, STAIR-actions の動作をほぼ識別できていないことがわかる. 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.

これらの結果から, 現段階では作成した合成データと現実のデータとのギャップは大きく, 現段階のようなシンプルな環境で動作をするような合成動画像では現実のデータの解析に使えないことがわかった. ここで, 作成した合成データと STAIR-actions との大きな違いとして, STAIR-actins の動画像においては対象が大きく映っているのに対し, 合成動画像の方では小さく映っているという写り具合の違いがある. そのため, 本データセットの今後の拡張として, 対象が大きく映るよう変更することが有効である可能性が考えられる.

5.3　実　験　2

本実験では, 作成した内装 (b) の合成動画像で事前学習を行ったモデルに, STAIR-actions でファインチューニングを行い, STAIR-actions でテストし評価する. この時, 照明条件のランダム化とノイズ・ぼかし処理による学習パフォーマンスの変化

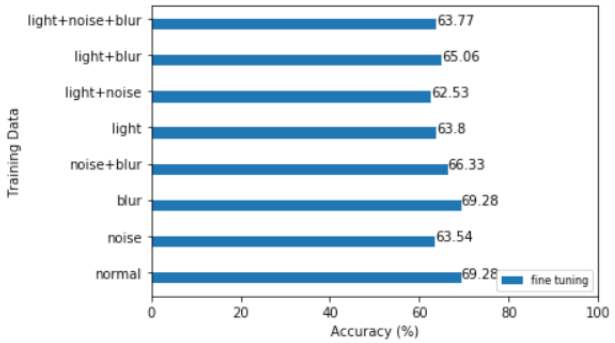

図 9　STAIR-actions によるファインチューニング後の動作識別の精度

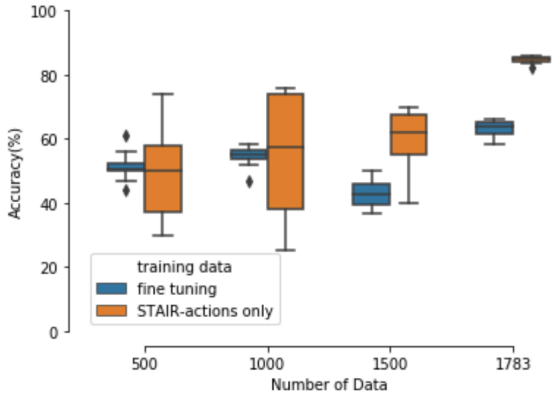

図 10　データサイズと精度の比較

を確かめるために, それぞれの有無を変更したデータで事前学習を行い, 実写動画像の動作識別への影響を調査する. さらに, 使用する STAIR-actions の動画像データの量を変化させ, 合成データで事前学習を行い STAIR-actions データでファインチューニングする場合と, STAIR-actions のみで学習を行う場合とでテスト精度を比較する.

データ数 1783 でのファインチューニングの精度を図 9 に示す. これより, ファインチューニングを行なうことで約 60-70%の精度で識別できるようになったことがわかる. また, 照明条件の変化・ノイズ・ぼかしによる影響は確認できなかった. これらの結果から, 合成動画像内の動作の特徴と, 現実の動画像での動作の特徴は, ある程度は共通しているということがわかる.

さらに, データ量を変化させた結果を図 10 に示す. 図 10 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な精度であり, 合成動画像を用いた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.

6.　まとめと今後の課題

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. 作成した合成データを用いた予備実験から, 作成したデータにおいて動作識別ができること, カメラ数が少なくデータ数が多い方が識別制度が高くなることがわかった. また, 実写データを用いた実験により, 作成した合成データでは現実のデータとのギャップを埋められていないが, それらのデータが持つ動作の特徴はある程度は共通していることがわかった. 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. さらに, 用意できる実写データが少ない場合は, 作成した合成データでの事前学習が有効であることを示した.

今後は実写画像を用いなくても十分に実写画像を識別できるようになることを目指し, 人部分が大きく映るようにカメラ環境を変更, また動作・人・背景を多様化して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.

文　　献

[1] Guangchun Cheng, Yiwen Wan, Abdullah N. Saudagar, Kamesh Namuduri, and Bill P. Buckles. Advances in human action recognition: A survey. *ArXiv*, abs/1501.05964, 2015.

[2] D. Wu, N. Sharma, and M. Blumenstein. Recent advances in video-based human action recognition using deep learning: A review. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2865–2872, May 2017.

[3] Chikako Takasaki, Atsuko Takefusa, Hidemoto Nakada, and Masato Oguch. A study of action recognition using pose data toward distributed processing over edge and cloud. *the 11th IEEE International Conference on Cloud Computing Technology and Science (CloudCom2019)*, pages pp.111–118, 2019.

[4] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

[5] 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.

[6] Joshua Tobin, Rachel H Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 23–30, 2017.

[7] Fereshteh Sadeghi and Sergey Levine. Cad^{2}rl: Real single-image flight without a single real image. *ArXiv*, abs/1611.04201, 2016.

[8] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtualworlds as proxy for multi-object tracking analysis. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4340–4349, 2016.

[9] Unity. `https://unity.com`.

[10] 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*, pages 6450–6459, 2017.

[11] Yuya Yoshikawa, Jiaqing Lin, and Akikazu Takeuchi. Stair actions: A video dataset of everyday home actions. *ArXiv*, abs/1804.04326, 2018.