3M2-4

家庭やオフィス内の動作認識用大規模動画データセットの構築

Constructing a Large-Scale Video Dataset for Human Action Recognition at Home and Office

吉川 友也　　竹内 彰一
Yuya Yoshikawa　Akikazu Takeuchi

千葉工業大学 人工知能・ソフトウェア技術研究センター
Software Technology and Artificial Intelligence Research Laboratory (STAIR Lab), Chiba Institute of Technology

This paper introduces a new large-scale video dataset for human action recognition at home and office, named STAIR Actions. At January, 2017, STAIR Actions contains 63,000 videos in total and 100 types of action labels. The length of most of the videos is five seconds, and each video corresponds to a single action label. In this paper, we explain how we construct STAIR Actions in details, along with showing the list of action labels and examples of the videos in STAIR Actions. Then, we show the result of human action recognition using STAIR Actions.

1. はじめに

人物動作認識 (human action recognition) は，与えられた動画中に映る人物が何の動作を行っているのか分類するタスクであり，映像解析における主要な研究テーマの一つとして盛んに研究されている [Marszałek 09, Heilbron 15, Sharma 15, Feichtenhofer 16].

近年の多くの人物動作認識の研究では，認識器としてディープニューラルネットワーク (DNN) に基づく手法が用いられる．このような手法では，大量の動作ラベル付き動画データを用いて DNN を学習し，未知の動画に対する動作ラベルを予測する．人物動作認識で使用される DNN は，画像認識で使われる 2 次元 Convolutional Neural Network (2D-CNN) をフレーム毎に適用して特徴量を抽出し，それを系列データに対して使われる Recurrent Neural Network (RNN) に入力して動作ラベルを出力するモデル [Donahue 15] や，2D-CNN を時間軸方向に拡張した 3D-CNN を使用するモデル [Ji 13] 等がある．

どのような DNN を使った場合でも，認識できる動作は学習で使用するデータセットに依存する．また，画像認識が大量の画像を学習データとして使用することによって成功したように，動画を入力とする人物動作認識も同様に，大量の動画を用いて DNN の学習を行うことが予測性能の向上に繋がる．従って，目的に合わせて認識したい動作を決定し，その動作が映る動画を大量に集めることが，人物動作認識の性能向上において重要である．

本研究では，人物動作認識技術のコミュニケーションロボットへの実装や介護老人保健施設での見守りシステムの構築を想定し，家庭やオフィスにおける人物動作認識に焦点を当て，大規模動画データセットの構築を行う．以降では，このデータセットのことを STAIR Actions と呼ぶ．2017 年 1 月時点で，STAIR Actions には 63,000 本の 5 秒動画が含まれている．各動画に対して，家庭やオフィスで見られる 100 種類の動作ラベルのうちの 1 つが付与されている．本論文では，どのように STAIR Actions を構築したのかを説明し，動作ラベルの一覧と，動作ラベル毎の動画の具体例を示す．また，STAIR Actions を用いた人物動作認識の実験結果を示す．

今後，STAIR Actions は，`http://actions.stair.center` からダウンロードできるように整備する．また，最終的には，各動作について 1,000 本の動画を用意し，合計 100,000 本の動画データセットとして公開する予定である．

連絡先: 吉川 友也，千葉県習志野市津田沼 2-17-1, yoshikawa@stair.center

表 1: STAIR Actions と代表的な人物動作認識データセットの比較.

データセット	動作の種類	全動画数
KTH [Schuldt 04]	6	600
Hollywood2 [Marszałek 09]	12	2,859
HMDB51 [Kuehne 11]	51	6,766
UCF-101 [Soomro 12]	101	13,320
ActivityNet 200 [Heilbron 15]	200	23,064
STAIR Actions	100	63,000

2. 関連研究

この節では，人物動作認識データセットの関連研究を紹介する．これまでに数多くの人物動作認識データセットが構築されている．表 1 は，代表的な人物動作認識データセットの動作の種類数と動画数を示す．KTH は，6 種類の動作の白黒動画のデータセットである．Hollywood2 は，69 本の映画から生成された短い動画のデータセットで，12 種類の動作ラベルと共に，10 種類のシーンラベルが付与されている．HMDB51 は，YouTube や映画等に対して 51 種類の動作ラベルを付与したデータセットである．UCF-101 は，YouTube の動画に対して 101 種類の動作ラベルを付与したデータセットで，約半数がスポーツに関する動作ラベルである．ActivityNet 200 は，YouTube の動画に対して 200 種類の動作ラベルが付与されており，現時点で最大の人物動作認識データセットである．

既存の人物動作認識データセットと比較して，STAIR Actions の主な特徴な 2 つある．1 つ目は，家庭やオフィス内の動作に特化している点である．HMDB51, UCF-101, ActivityNet 200 のような動作の種類が多いデータセットの場合，スポーツや家庭，アウトドアなど多様なジャンルの動作が含まれる．一方で，STAIR Actions では，家庭やオフィス内といった限られた範囲で見られる動作のみを対象としており，家庭やオフィスで人物動作認識を行いたい場合により適したデータセットとなっている．2 つ目は，我々の知る限り，現時点で最も動画数の多い人物動作認識のデータセットとなる点である．動画

表 2: STAIR Actions の動作ラベル一覧

赤ちゃんにミルクを与えている, コンタクトレンズを入れている・外している, 人の頭を撫でている, 盤ゲームをしている, じゃんけんをする, 顔を洗っている, ハイタッチする, うがいをしている, 赤ちゃんにおむつをかえている, 飛び跳ねている, 驚いている, 冷蔵庫のドアを開ける, おんぶしている, 赤ちゃんが這い這いしている, ゴミを捨てている, ネイルをしている, 幼児に食事を食べさせている, 握手する, 髭を剃っている, 泣いている, 洗濯物を干している・取り入れている, 喧嘩している, 食器を洗っている, ネクタイを締める, アイロンをしている, 部屋に入る, お茶をいれる, 車椅子で移動している, 手を洗っている, 怒っている, マッサージをしている, 倒れる・転ぶ, 渡す, テレビを見ている, 赤ちゃんが泣いている, 家電製品をリモコンで操作をする, ハグをする, 靴を磨いている, 靴を履く, 部屋から出る, キスをする, ゲーム機で遊んでいる, 窓を拭いている, 洗濯物をたたんでいる, 階段を上っている・降りている, 歯を磨いている, 音楽を聴いている, タバコを吸っている, お辞儀する, 髪を乾かしている, 裁縫をしている, 勉強している, 新聞を読んでいる, 写真を撮る, 寝ている, 掃除をしている, 椅子・机・脚立に乗る, 床に横たわっている, メガネをかける, 庭仕事している, 文字を書いている, 服を脱ぐ, 走り回っている, 編み物をしている・刺繍をしている, エクササイズをしている, 容器を開閉する, 電話をしている, 服を着る, ご飯を食べている, 拍手している, 座る, 抱っこしている, 投げる, 遊んでいる, 紙を折っている, 本を読んでいる, PC を使っている, 飲んでいる, 食べている, 立ち上がる, 食材を切っている, 料理を作っている, 髪をセットしている・髪をブラッシングしている, 化粧をしている・口紅を塗っている, スマホを操作している, タブレットを使っている, ギターを弾いている, ピアノを弾いている, 笛を吹いている, 絵を描いている, 教えている, 頷く, 話している, 人の話を聞いている, 指をさす, 動物をなでる, 歩いている, 踊っている, 喜んでいる・笑っている, 首を横に振る

数が多くなればなるほど，様々なシーンや角度からの動画が含まれるようになり，人物動作認識の性能向上が期待される．

3. STAIR Actions

STAIR Actions は，家庭やオフィスで見られる人の動作を認識するための動画データセットである．2017 年 1 月時点で，STAIR Actions には 63,000 本の 5 秒動画が含まれている．また，各 5 秒動画に対して，100 種類の動作ラベルのうちの 1 つが付与されている．表 2 は，STAIR Actions に含まれる 100 種類の動作ラベルの一覧を示す．また，図 1 では，各動作ラベルに対応する動画の具体例を示す．

以下では，STAIR Actions の構築方法について説明する．STAIR Actions は，(1) YouTube 上の動画と，(2) STAIR Actions の構築のために撮影された動画で構成される．まず，(1) YouTube 上の動画については，以下のステップで動作ラベルを付与した．

ステップ 1. YouTube から動画を取得 (3.1 節)

ステップ 2. 動画から 5 秒動画の切り出し (3.2 節)

ステップ 3. 5 秒動画に対してクラウドソーシングでラベル付け (3.3 節)

ステップ 4. ラベル付け結果の検品 (3.4 節)

以下の 3.1–3.4 節では，それぞれのステップを詳細に説明する．また，(2) STAIR Actions の構築のために撮影された動画に関しては，3.5 節で説明する．

3.1 YouTube から動画を取得

動作ラベルを付与する動画は，YouTube から取得した．動画は，動作ラベルに関連するキーワードを自作し，それを用いて検索して取得した．検索の際には，長さが 4 分未満で，Creative Commons を主張する動画が見つかるように設定した．

3.2 動画から 5 秒動画の切り出し

YouTube から取得した動画の長さは，数秒から数分まで様々である．しかし，動画中の全ての時間で一つの動作が行われていることはほとんどなく，その動作は多くの場合数秒で終了する．また，クラウドソーシングでラベル付けを行うことを考慮すると，長い動画を作業者に見せて，いつ何の動作が行われているのかをラベル付けしてもらう作業は複雑で非効率である．そこで，取得した動画を 5 秒毎に切り出し，5 秒動画を作成した．

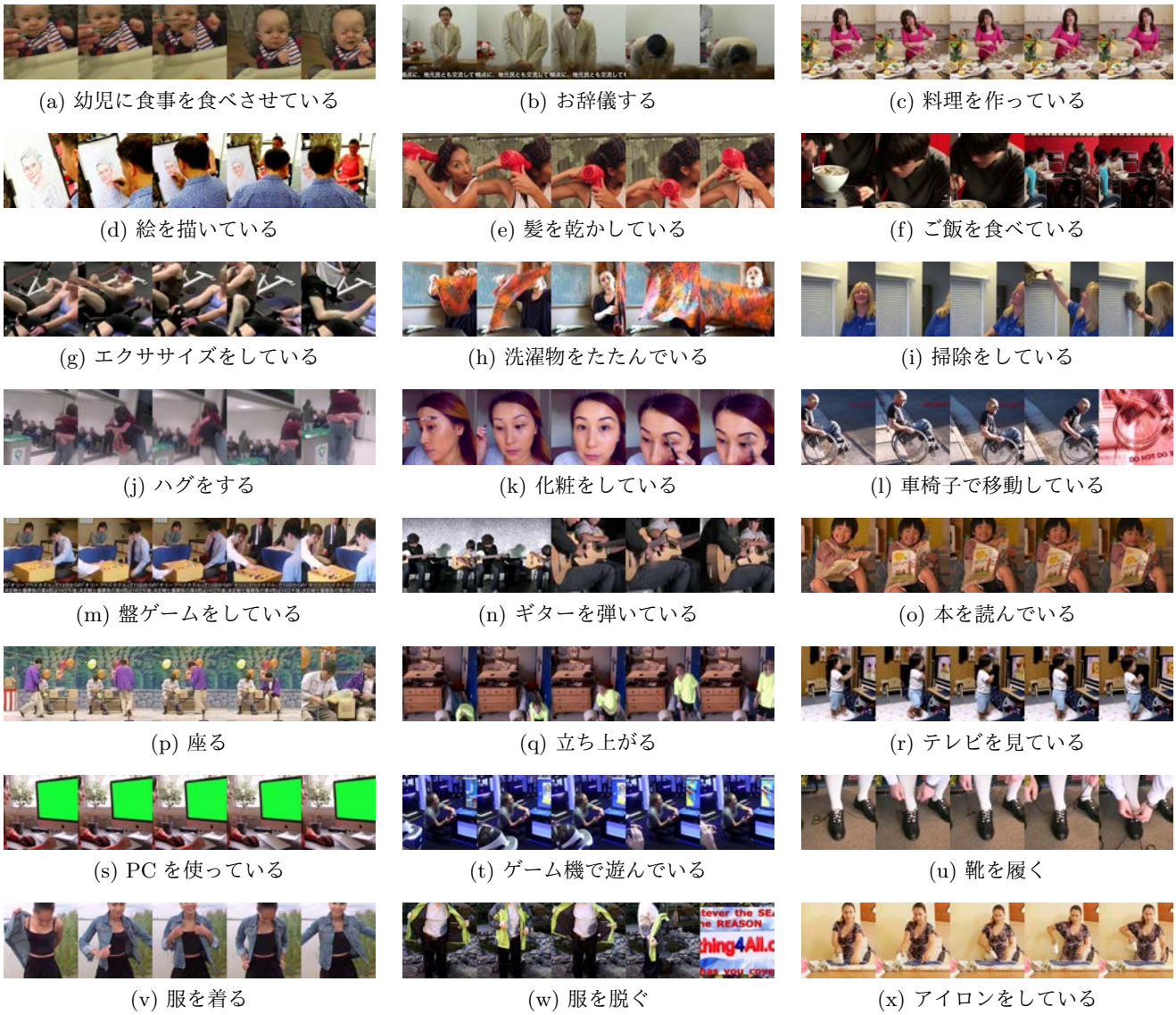
その上で，今回は人物動作認識のためのデータセットを作成するため，人が映っていない 5 秒動画はクラウドソーシングでラベル付け作業に入る前になるべく排除したい．そこで，まず最初に，写真に音楽が付いているだけの静止動画や，アニメやゲーム画面が映った動画を排除する前処理を行った．その後で，各 5 秒動画に対してフレーム毎に人検出を行い，人が検出された 5 秒動画のみをラベル付け作業の対象とした．

3.3 5 秒動画に対してクラウドソーシングでラベル付け

5 秒動画に対するラベル付け作業は，クラウドソーシングを利用して行った．ラベル付け作業者には，事前にラベル付けのガイドラインを示した．また，不適切なラベルを大量につける悪意のある作業者を排除するために，ガイドラインの理解度テストを行い，そのテストに合格した人にのみラベル付け作業を依頼するようにした．

クラウドソーシングでのラベル付け作業を効率的に行うために，独自のラベル付け Web システムを構築した．図 2 は，ラベル付け Web システムの作業画面を示す．この Web システムでは，作業者は 5 秒動画を見て，10 種類の動作ラベルと「該当なし」ラベルの計 11 個の選択肢から 1 つを選択する．10 種類の動作ラベルのみ表示する理由は，100 種類の選択肢から 1 つ選ぶようにすると，作業者が該当する選択肢を探すだけで時間が掛かってしまい非効率なためである．なお，この 10 種類の動作ラベルを 1 つの動作ラベルグループとして設定し，作業中にこの動作ラベルグループが変更されることはない．100 種類の動作ラベルを網羅するためには，同じ 5 秒動画に対して 10 回のラベル付け作業が必要である．しかし，一度 5 秒動画にラベルが付与されると，次から同じ動画が表示されることのないようにしているため，実際には同じ 5 秒動画は平均 5 回表示された．

質の良いラベルを得るためには，作業者にガイドラインを理解しラベル付けを行ってもらう必要がある．しかし，各動作についてラベル付けのガイドラインがあるため，100 種類の動作全てのガイドラインを作業者が覚えることは困難である．したがって，作業者を 1 つの動作ラベルグループ (10 種類の動作ラベル) に対応付け，その作業者が覚える必要のあるガイド



(a) 幼児に食事を食べさせている　(b) お辞儀する　(c) 料理を作っている

(d) 絵を描いている　(e) 髪を乾かしている　(f) ご飯を食べている

(g) エクササイズをしている　(h) 洗濯物をたたんでいる　(i) 掃除をしている

(j) ハグをする　(k) 化粧をしている　(l) 車椅子で移動している

(m) 盤ゲームをしている　(n) ギターを弾いている　(o) 本を読んでいる

(p) 座る　(q) 立ち上がる　(r) テレビを見ている

(s) PC を使っている　(t) ゲーム機で遊んでいる　(u) 靴を履く

(v) 服を着る　(w) 服を脱ぐ　(x) アイロンをしている

図 1: 動作ラベル毎の動画の例．動画から 1 秒毎にサムネイルを生成し横に並べて表示した．

ラインが少なくなるようにした．

3.4　ラベル付け結果の検品

クラウドソーシングで得られたラベルの品質を保証するために，前ステップで作られた全てのラベル付き 5 秒動画に対して，正しくラベル付けされているかどうかの検品を行った．検品もまたクラウドソーシングを用いて行った．検品作業では，各 5 秒動画は，3 人の検品作業者によって正しいラベルが付けられているか確認した．その上で，少なくとも 2 人の検品作業者が正しいラベルであると判断した 5 秒動画のみを STAIR Actions に含めた．

3.5　動作動画の撮影

この節では，STAIR Actions の構築のために撮影された動画に関して説明する．動作動画の撮影は，クラウドソーシングを利用して行った．具体的には，表 2 の動作を行った 5 秒程度の動画を作業者の所有するカメラで撮影し提出してもらうように依頼した．なお，他人の動画を許可なく撮影して提出されることを避けるため，撮影時には，“Stair Lab.” と記入した紙を掲示するように作業者に指示した．図 3 は，動作動画の撮影で得られた「部屋から出る」ラベルの動画の例である．提出された全ての動画は，検品者によって動作ラベルに適合した動画になっているかを確認し，適合した動画のみを STAIR Actions に含めた．

4.　人物動作認識の実験

この節では，STAIR Actions を用いた人物動作認識の実験結果を示す．

人物動作認識には，[Donahue 15] と同等の 2D-CNN と LSTM を組み合わせた DNN を使用した．2D-CNN の部分には ImageNet を使用して事前学習した AlexNet [Krizhevsky 12] を使用した．

表 3 は，我々が実装した DNN による人物動作認識の精度を示す．STAIR Actions における認識精度に加えて，UCF-101 と HMDB51 における認識精度も参考のために併記した．UCF-101 や HMDB51 と比較して，STAIR Actions の認識精度が低い理由は，動作ラベルの粒度の細かさが原因だと考えられる．UCF-101 と HMDB51 は共に，スポーツに関する動作

図 2: ラベル付け Web システムの作業画面

図 3: 動作動画の撮影で得られた「部屋から出る」ラベルの動画

や家庭内で見られる動作等，様々な動作に関する動画が含まれている．そのような場合は，データセットの中で似ている動作が少なく，結果的に認識しやすくなると考えられる．一方で，STAIR Actions は，家庭やオフィス内の動作に絞っているため，データセット内で似ている動作が多い．そのため，他の 2 つのデータセットに比べて精度が低くなったと考えられる．

5. おわりに

本論文では，我々が新たに構築した，家庭やオフィス内で見られる動作 100 種類を認識するための大規模動画データセット STAIR Actions を紹介した．2017 年 1 月時点で，STAIR Actions は 63,000 本の動画から構成され，各動画には 1 つの動作ラベルが付与されている．最終的には，各動作について 1,000 本の動画を用意し，合計 100,000 本の動画データセットとして公開する予定である．

謝辞

この成果は，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．

データセット構築に協力していただいた株式会社 mokha 蒲地氏，スケールアウト株式会社 荻野氏，タノシム株式会社 堤氏に感謝致します．

表 3: 人物動作認識の精度

STAIR Actions	UCF-101	HMDB51
45.60%	86.04%	52.87%

参考文献

[Donahue 15] Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T., Austin, U. T., Lowell, U., and Berkeley, U. C.: Long-term Recurrent Convolutional Networks for Visual Recognition and Description, in *IEEE Conference on Computer Vision and Pattern Recognition* (2015)

[Feichtenhofer 16] Feichtenhofer, C., Pinz, A., and Zisserman, A.: Convolutional Two-Stream Network Fusion for Video Action Recognition, in *IEEE Conference on Computer Vision and Pattern Recognition*, No. i, pp. 1933–1941 (2016)

[Heilbron 15] Heilbron, F. C., Escorcia, V., Ghanem, B., and Niebles, J. C.: ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding, in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 961–970, IEEE (2015)

[Ji 13] Ji, S., Yang, M., Yu, K., and Xu, W.: 3D Convolutional Neural Networks for Human Action Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 1, pp. 221–31 (2013)

[Krizhevsky 12] Krizhevsky, A., Sutskever, I., and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, *Advances In Neural Information Processing Systems*, pp. 1–9 (2012)

[Kuehne 11] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T.: HMDB: A Large Video Database for Human Motion Recognition, in *International Conference on Computer Vision*, pp. 2556–2563, IEEE (2011)

[Marszałek 09] Marszałek, M., Laptev, I., and Schmid, C.: Actions in Context, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, No. i, pp. 2929–2936 (2009)

[Schuldt 04] Schuldt, C., Laptev, I., and Caputo, B.: Recognizing Human Actions: A Local SVM Approach, in *17th International Conference on Pattern Recognition*, pp. 32–36, IEEE Computer Society (2004)

[Sharma 15] Sharma, S., Kiros, R., and Salakhutdinov, R.: Action Recognition using Visual Attention, *arXiv preprint*, pp. 1–11 (2015)

[Soomro 12] Soomro, K., Zamir, A. R., and Shah, M.: UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild, Technical Report November (2012)