Home Action Genome：対照学習を用いた階層的行動認識

Home Action Genome: Cooperative Compositional Action Understanding

小塚 和紀　Kazuki Kozuka
石坂 隼　Shun Ishizaka

要　旨

本研究では，人が物事を理解する際に重要な要素である，階層構造や複数のモダリティを収録した，複数視点で撮影された新たな行動認識データセットであるHome Action Genome（HOMAGE）を提案する．複数のモーダルおよび複数の視点の情報を利用し，対照学習を用いた階層的な行動認識のための学習フレームワークCooperative Compositional Action Understanding（CCAU）により，検証したすべてのモダリティにおいて一貫した性能向上を示した．

Abstract

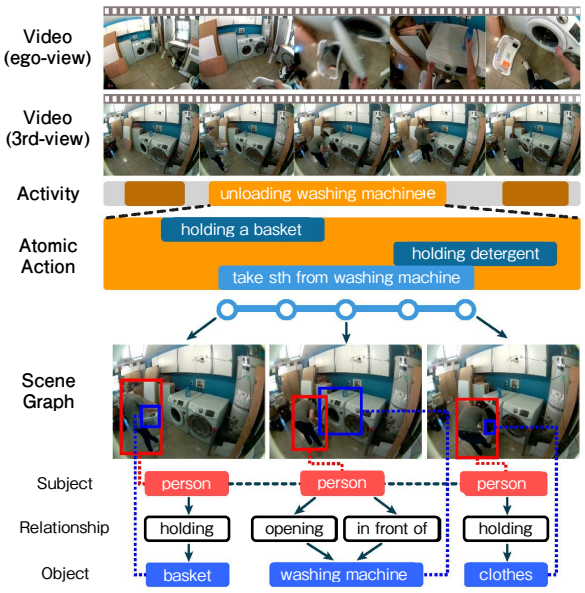
In this research, we propose the Home Action Genome (HOMAGE), a new action recognition dataset containing hierarchical structures and multiple modalities captured from multiple viewpoints, which is an important factor for human to understand things. We show that Cooperative Compositional Action Understanding (CCAU), a learning framework for hierarchical action recognition using contrast learning, consistently improves performance across all modalities tested.

1. はじめに

動画中の行動を認識することは，実世界でのさまざまなアプリケーションへの応用が可能であり，非常に重要な課題である．しかし近年，画像分類などの機械学習の手法が非常に高い精度で行えるようになったにも関わらず，例えば住空間での人物の行動のような複雑な行動やイベントを認識することにはいまだに課題が多く残っている．

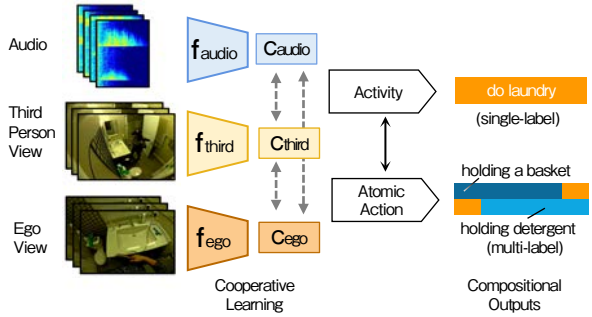
これまで，さまざまな行動認識向けデータセットが提案されている．ActivityNet [1]やKinetics [2], Charades [3], UCF101 [4]は，映像中の一般的な行動認識や行動区間の特定などを目的として，YouTubeなどの動画を切り抜いて構築されている．また，EPIC Kitchens [5]は，一人称視点の映像における認識課題を解くためのものである．Action Genome [6]は物体・人物間の関係性情報を行動認識に活用したものである．他にも，LEMMA [7]のように，複数の視点から行動を観察し，認識することに焦点を当てたデータセットが存在する．一方で，対照学習[8]などの，マルチモーダル・マルチビューのデータセットに適用可能な手法が近年提案されており，物体認識やポーズ認識などの個々の要素そのものの認識精度は向上している．しかし行動は物体の性質や多数の物体間の関係性などから構成されており，このようなモデルでの行動認識はいまだ難しい課題である．Sayed [9], Simonyan [10], Tian [11]のように複数のモーダルを用いることは有用であるが，これらの課題を統合するベンチマークは存在しない．これらを踏まえ，本論文では，階層構造に注目した行動認識の新しい手法，課題に取り組むためのデータセットであるHome Action Genome（HOMAGE）を提案する．

HOMAGEには，**第1図**に示す時間同期されている複数の視点で撮影されたマルチモーダル情報と，階層的な行動ラベル・詳細行動ラベルが含まれている．家庭内での行動では，物体による遮蔽や，長時間の行動，物体との相互作用などを扱う必要があり，単一の視点の映像だけでこれらを扱うのは難しい．本データセットでは，**第2図**に示す複数の視点，モーダルで行動を撮影することで物体の遮蔽に対応し，またシーングラフ情報で物体間の関係性を把握することが可能である．さらに，**第3図**に示す映像・音声・シーン構成情報といった複数のモダリティを用いて学習を行う新しい手法をベンチマークとして提案する．学習においては，表現空間を構築するためにすべての視点映像とモダ



第1図　HOMAGEにおけるラベル情報
Fig. 1　Label information in HOMAGE

第2図　HOMAGEの映像例
Fig. 2　Examples of HOMAGE videos



第3図　提案手法の概要図
Fig. 3　Overview of the proposed method

リティからの情報を活用している．提案手法をHOMAGEで学習することにより，推論時には他のモダリティを用いなくても，性能が向上することを示した．

本研究での貢献を要約すると，以下のとおりである．

- 複数視点・モダリティをもち，時空間シーングラフと行動の階層構造を付与した新しいデータセットHome Action Genome（HOMAGE）を構築した．
- 複数のモダリティと階層的な行動ラベルを活用することで，個々のモダリティのみで学習した手法よりも高性能な学習フレームワークCooperative Compositional Action Understanding（CCAU）を提案した．

2. 関連研究

関連データセット　これまでに複数のモーダルが収録されている行動認識データセットは幾つか提案されている．NTU RGB+D [12]はRGB・深度・赤外線情報を80のシーンで撮影し，人物の骨格の3次元情報を付与している．しかし，行動ラベルについては各映像に単一のものが付与されているだけであり，行動区間の推定タスクには利用できない．MMAct [13]はRGB・骨格情報・加速度・角速度・方位情報といったマルチモーダルデータを収録した行動認識データセットである．これは一人称視点映像と4つの三人称視点映像から構成されており，行動区間の情報が付与されている．しかし，MMActでは物体の位置や物体間の関係性の情報は付与されてない．LEMMA [7]は，マルチビュー・マルチエージェントの行動認識データセットである，三人称視点映像において物体のバウンディングボックスが付与されており，また行動ラベルと，動詞・名詞の形で付与された行動の構成情報を含んでいる．しかし，これらのデータは人物と関わる物体の領域情報を含んでいない．Action Genome [6]は，Charades [10]の映像に時空間シーングラフラベルの情報を付与したものであるが，単一視点の映像しか収録していない．これら既存のデータセットと比較して，HOMAGEは，**第2図**に示すような複数視点で撮影された行動の映像に対して，マルチモーダル・マルチビューの情報，詳細行動とその行動区間，人物とその人物に関わる物体のバウンディングボックスおよび関係性を記述した時空間シーングラフを含んでいる．

マルチモーダル学習　複数のモダリティは認識において重要な情報源である．[11], [14]では，自己教師有り学習の枠組みで，モダリティ間の相互情報を最大化する対照学習の手法が提案されている．これらの手法は任意の数のモダリティに拡張することができる．行動認識においても複数のモダリティを用いることは重要であり，[10]では，オプティカルフローを利用することで性能が大幅に向上することが示されている．

3. Home Action Genome

HOMAGE（Home Action Genome）は，一人称視点を含む複数の視点からの，マルチモーダルな時間同期された住空間での行動の映像情報に，詳細行動ラベルおよび時空間シーングラフを付与した，行動認識のための新しいベンチマークである．本章では，HOMAGEのデータセットの設計やアノテーションの枠組みについて説明する．

3.1　データ収集

27人の参加者が，2つの家のキッチン・バスルーム・寝室・リビング・ランドリールームで行動する様子を撮影した．センサとしては，カメラ（RGB）・赤外線（IR）・マイク・照度（RGB／高感度）・人感・加速度・角速度・地磁気・気圧・湿度・温度の12種類を使用した．センサは部屋の複数個所および参加者の頭部に設置した．また，すべての視点のセンサは時間同期している．以下では，これらの設定で，特定の行動から収集されるデータのセットを行動シーケンスと呼ぶことにする．

3.2　アノテーションの枠組み

Home Action Genomeには70種類の行動ラベルと，453種類

の詳細行動ラベルが付与されている．また各詳細行動区間に基づいて，物体領域および人物と物体間の関係性ラベルが付与されている．物体領域についてはバウンディングボックスにて与えられ，各領域に対し82種類の物体種別ラベルを振っている．なお，行動ラベルは住空間での行動を網羅するために，American Time Use Survey（ATUS）[15]から選定した. ATUSは日常行動における行動を費やす時間をもとに整理したものであり，このリストから宅内における行動のなかで高頻度かつ長時間行われるものを選定した．これにより，本データセットは宅内における行動を広範囲で網羅できる．詳細行動ラベルについては，選定された70種類の行動ラベルに基づく行動を計測した結果から決定した.

これらのアノテーションを撮影された各動画に付与する枠組みを**第1図**に示す. まず動画に行動ラベル付与し，さらに，各行動内での詳細行動ラベルおよびその行動区間を付与する．例えば，行動「unload washing machine」に対する詳細行動は，その行動の途中で行われる基本動作にあたる「holding a basket」などが該当する．また，行動区間は同一行動が継続している区間の情報であり，行動の開始，終了時刻を付与する. 各詳細行動区間のなかから等間隔に3フレームあるいは5フレームを抽出して, それらのフレームに対して物体領域ラベルおよび人物と物体の関係性ラベルを付与する．なお行動区間が2秒以下の際には3フレーム，それより長い場合は5フレームを抽出した. このように行動に基づいてラベルを付与することで，各詳細動作に対して，位置と関係性の大局的な変化を捉えることができる.

3.3 データセットの統計

HOMAGEには1752の行動シーケンス，5700の映像に対し，70種類の行動ラベルと453の詳細行動をアノテーションした．詳細行動としては，24569個の詳細行動区間が付与されている．時空間シーングラフとしては，各行動シーケンスのうち1つの三人称視点映像に, 物体と人物のバウンディングボックスおよび関係ラベルを付与している. HOMAGEにはpersonを除く86種類の物体クラスと, 29の関係性クラスが含まれており，その数はバウンディングボックスが497534個，関係性ラベルが583 481個である．**第1表**にHOMAGEと既存の行動認識向けデータセットの比較表を示す．**第1表**に示すように，本データセットは複数の視点・モーダルとともに，物体との相互作用などを扱うシーングラフを保有する世界初のデータセットであり，これにより住空間における複雑な行動認識の新たな手法，課題に取り組むことが可能となる.

4. CCAU

HOMAGE における豊富なアノテーションを活用し，行動認識の性能を向上させる手法として，Cooperative Compositional Action Understanding（CCAU）を提案する．**第3図**に示すように, CCAUでは複数のモダリティによる協調学習を同時に行うことで，行動およびそれに関連する詳細行動の精度を向上させる．本章ではこの具体的な手法について説明する．本論文では，モダリティとして複数の映像視点・音声・シーングラフを利用する.

4.1 マルチモーダル協調学習

フレーム長T，解像度$H\times W$，C個のチャンネルをもつ動画Vを$\{\mathbf{i}_1, \mathbf{i}_1, \ldots, \mathbf{i}_T\}$ ($\mathbf{i}_t\in\mathbb{R}^{H\times W\times C}$)と表現する．ある映像$V$を$N$個のブロック$V=\{x_1, x_2, ..., x_N\}$に分割する．ここで，Kはブロックあたりのフレーム数であり，このとき，$x_j\in\mathbb{R}^{K\times H\times W\times C}$となる．各入力ブロック$x_j$を潜在表現$z_j$に変換するエンコーダを$f(\cdot)$とし，$f(\cdot)$により得られる潜在表現のシーケンスを入力として，コンテキスト表現c_jを生成する集約関数を$g(\cdot)$とする．すなわち，$c_j=g(z_1, z_2, ..., z_j)$である．本手法では，$z_j\in\mathbb{R}^{H'\times W'\times D}$かつ$c_j\in\mathbb{R}^{\mathrm{D}}$としている．ここで，$D$は埋め込みサイズであり，$H'$, W'はダウンサンプルした解像度を指す．なお本研究では

第1表　HOMAGEと既存の行動認識向けデータセットの比較
（Seq: 収録動画数，HL: 高次元の行動ラベル，TL: 詳細行動ラベル，SG: シーングラフ）

Table 1　HOMAGE and existing datasets for action recognition
(Seq: number of videos, HL: high-dimensional action labels, TL: detailed action labels, SG: scene graphs)

Dataset	Seq	hrs	Modalities	Views	HL	HL Classes	TL	TL Classes	TL Ins	SG
UCF101 [4]	13K	27	1	1	✓	101	-	-	-	-
ActivityNet [1]	28K	648	1	1	✓	200	-	-	-	-
Kinetics-700 [2]	650K	1.79K	1	1	✓	700	-	-	-	-
AVA [16]	430	108	1	1	-	-	✓	80	1.58M	-
EPIC-Kitchen [5]	-	55	1	1	-	-	✓	125	39.6K	-
MMAct [13]	36K	-	6	5	-	-	✓	37	36.8K	-
Action Genome [6]	10K	82	1	1	-	-	✓	157	66.5K	✓
LEMMA [7]	324	10.1	2	4	✓	15	✓	863	11.8K	-
Ours	1.75K	25.4	12	2〜5	✓	75	✓	453	24.6K	✓

H'=4, W'=4, D=256とした．この一連の変換を$c=F(V)(F(\cdot)=g(f(\cdot)))$と定義する．算出された表現$c$は，ブロックごとに分類に利用され，行動ラベルや詳細行動ラベルなどの予測値を生成する．複数のモダリティの場合，$c_m=F_m(V_m)$のように拡張する．ここで，V_m, c_mおよびF_mはそれぞれ，入力するモーダル，コンテキスト表現およびモダリティmに関するエンコーダを指す．

本手法では各モダリティのエンコーダを同時に学習することで，モダリティ同士で補完的に情報を与え合うことを期待している．この相補的な学習を促進するため，CCAUでは対照的なマルチモーダル損失関数を利用している．ここでは，異なるモダリティm, m'のアライメントを行うことを考えるため，[11] と同様に，（1）式で表されるコンテキスト表現cに対するNCE損失関数を使用する．これは，動画に含まれるすべての負のペアから，1つの正のペアを識別するクロスエントロピー損失である．

$$\mathcal{L}_{align}^{m,m'} = -\sum_i (log \frac{\exp(c_i^m \cdot c_i^{m'})}{\sum_j \exp(c_i^m \cdot c_j^{m'})} \quad \cdots\cdots\cdots (1)$$

これを複数視点に拡張するには，すべてのペアに対して同じ損失関数を使用し，同時に最適化すればよい，つまり$\mathcal{L}_{align} = \sum_{m,m'} \mathcal{L}_{align}^{m,m'}$と表現できる．

4.2 階層構造を利用した行動認識

HOMAGEではマルチモーダルの情報に加えて，行動ラベルと，それを構成する詳細行動ラベルを利用することができる．詳細行動ラベルを利用することは，全体的な行動の認識だけでなく，行動と詳細行動との関係性の学習にも有用である．

CCAUでは行動ラベルと詳細行動ラベルを同時に利用することで，行動の階層構造を学習し，認識精度を向上させている．映像レベルの行動クラスと詳細行動クラスの双方を予測するために，コンテキスト表現c を活用する．ここで，行動クラスの認識は標準的なワンショット分類タスクとして，詳細行動クラスの認識はマルチターゲット分類タスクとして定式化する．それぞれの損失関数を$\mathcal{L}_{video} = \mathcal{L}_v$, $\mathcal{L}_{atomic} = \mathcal{L}_a$としたとき，階層構造全体の損失は$\mathcal{L}_{composition} = \mathcal{L}_c = \mathcal{L}_v + \lambda\mathcal{L}_a$と表せる．

5. 実験

協調学習と行動の構造を学習する枠組みにより，CCAUはHOMAGEに含まれる豊富な情報を利用し，行動認識の性能を改善した．本章では，CCAUの有用性を検証するための実験について説明する．

5.1 実装詳細

すべての実験について，各視点映像はそれぞれ別のモダリティとして扱う．画像については，各入力フレームを128x128のサイズにリサイズする．エンコーダf(·)としては3D-ResNetを利用した．実験には3DResNet18を使用し，以下ではResNet18と表記する．集約関数$g(\cdot)$としてはカーネルサイズ(1,1)の1層のConvolutional Gated Recurrent Unit（ConvGRU）を用いた．重みは特徴量マップのすべての空間位置で共有されるように設計されているため，集約関数が時間軸方向に特徴を伝播（でんぱ）させることができる．音声については，まず音声をMP3形式に変換し，log-melスペクトルを計算したのち，畳み込みアーキテクチャ（VGG19）に通す構造にした．

また，学習時にはモデルは他のモーダルへアクセスすることが可能であるが，推論時には単一のモーダルのみを使用する．そのため，評価は個々のモダリティのみを用いた推論を対象とした．

5.2 結果

〔1〕複数モーダルを用いた学習の有用性

複数のモダリティを用いた学習が性能向上に寄与することを検証するため，(1) 各モダリティを個別に学習する手法，(2) 一人称視点映像を含む，複数の視点映像を同時に学習に用いる手法，(3) 複数視点の音声情報を学習に用いる手法について性能を評価する．動画単位での行動認識の結果を**第2表**に示す．複数モーダルを用いて学習させることで，認識の性能が向上しており，CCAUがモーダル間に存在する相補的な情報を活用できているといえる．

第2表　動画単位での行動認識の結果
Table 2　Results of per-video recognition

Method	Audio	Ego	3rd
(1) Single Modality	28.5	31.3	21.8
(2) Coop - Ego +3rd	-	35.1	23.5
(3) Coop – Ego + 3rd + Aud	**33.3**	**37.7**	**24.7**

〔2〕階層構造学習の有用性

CCAUは，行動と詳細行動の両方を学習に用いることで，階層構造を学習し，性能が向上するという仮説に基づいている．この仮説を検証するために，(1) 行動ラベルで学習，(2) 詳細行動ラベルで学習，(3) 両方のラベルを協調させずに学習，(4) 両方のラベルを協調して学習，とした場合での性能を比較した．**第3表**のとおり，(4) の場合に行動の認識精度と詳細行動の認識精度の双方が向上した．

第3表　階層構造学習の結果
Table 3　Results of Compositional learning

Method	Acc (Activity)			mAP (Atomic Action)		
	Audio	Ego	3rd	Audio	Ego	3rd
(1)	28.3	31.1	17.0	-	-	-
(2)	-	-	-	5.9	18.5	9.5
(3)	23.5	32.1	16.2	16.4	26.3	12.2
(4)	**29.3**	**34.9**	**19.2**	**21.7**	**29.3**	**13.8**

6. まとめ

複数の視点・モダリティをもち，階層化された行動および詳細行動ラベルをもつ行動認識データセットであるHome Action Genome（HOMAGE）を構築した．また，複数のモダリティの情報とHOMAGEにおける行動の階層構造を利用し，より良い表現を学習する手法としてCCAUを提案した．本論文では，行動の階層構造を学習することで，認識性能を向上できることを示した．HOMAGEはマルチモーダルデータと階層的な情報をもつため，時空間シーングラフを用いた推論や説明可能な行動認識，プライバシを考慮した認識タスクなどに利用することが期待される．

参 考 文 献

[1] Fabian Caba Heilbron et al., ActivityNet: “A Large-Scale Video Benchmark for Human Activity Understanding,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, pp. 961–970, 2015.

[2] Carreira, J et al., “A short note on the kinetics-700 human action dataset,” arXiv preprint arXiv:1907.06987, 2019.

[3] Sigurdsson, G. A. et al., Hollywood in homes: “Crowdsourcing data collection for activity understanding,” Proceedings of the European Conference on Computer Vision (ECCV), Springer, Amsterdam, pp. 510–526, 2016.

[4] Soomro, K. et al., “UCF101: A dataset of 101 human actions classes from videos in the wild,” arXiv preprint arXiv:1212.0402, 2012.

[5] Damen, D. et al., “Scaling egocentric vision: The epic-kitchens dataset,” Proceedings of the European Conference on Computer Vision (ECCV), Munich, pp. 720–736, 2018.

[6] Ji, J. et al., “Action Genome: Actions as Compositions of Spatio-Temporal Scene Graphs,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10236–10247, 2020.

[7] Jia, B. et al., “A Multi-view Dataset for Learning Multi-agent Multi-task Activities,” Proceedings of the European Conference on Computer Vision (ECCV), pp. 767–786, 2020.

[8] Chen, T et al., “A simple framework for contrastive learning of visual representations, International conference on machine learning,” PMLR, pp. 1597–1607, 2020.

[9] Sayed, N. et al. “Cross and Learn: Cross-Modal Self-Supervision,” CoRR, abs/1811.03879, 2018.

[10] Simonyan, K. et al., “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, MIT Press, Massachusetts, pp. 568–576, 2014.

[11] Tian, Y. et al., “Contrastive Multiview Coding, “CoRR, abs/1906.05849, 2019.

[12] Shahroudy, A. et al., “Ntu rgb+ d: A large scale dataset for 3d human activity analysis,” Proceedings of the IEEE conference on computer vision and pattern recognition,, Las Vegas, pp. 1010–1019, 2016.

[13] Kong, Q. et al., “MMAct: A Large-Scale Dataset for Cross Modal Human Action Understanding,” Proceedings of the IEEE International Conference on Computer Vision, Seoul, pp. 8658–8667 2019.

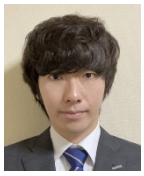
[14] Korbar et al., “Cooperative learning of audio and video models from self-supervised synchronization,” arXiv preprint arXiv:1807.00230, 2018.

[15] Hamermesh, D. S.et al., “Data watch: The American time use survey,” Journal of Economic Perspectives, vol. 19, no. 1, pp. 221–232, 2005.

[16] Gu, C. et al., “Ava: A video dataset of spatiotemporally localized atomic visual actions,” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, pp. 6047–6056, 2018.

執筆者紹介

小塚　和紀　　Kazuki Kozuka
テクノロジー本部　デジタル・AI技術センター
Digital & AI Technology Center, Technology Div.
博士（工学）



石坂　隼　　Shun Ishizaka
テクノロジー本部　デジタル・AI技術センター
Digital & AI Technology Center, Technology Div.