The 38th Annual Conference of the Japanese Society for Artificial Intelligence, 2024

調理作業理解のための言語資源付き固定視点映像データセットの構築

Unedited Fixed-viewpoint Procedural Videos with Language Resources for Understanding Cooking Activities

橋本 敦史 *1 Atsushi Hashimoto
前田 航希 *1*2 Koki Maeda
平澤 寅庄 *1*3 Tosho Hirasawa
原島 純 *4 Jun Harashima
Leszek Rybicki*4 Leszek Rybicki
深澤 祐援 *4 Yusuke Fukasawa
牛久 祥孝 *1 Yoshitaka Ushiku

*1オムロンサイニックエックス株式会社 OMRON SINIC X Corp.
*2東京工業大学 Tokyo Institute of Technology
*3東京都立大学 Tokyo metropolitan university
*4クックパッド株式会社 Cookpad Inc.

Large-scale web videos have contributed significantly to recent progress in video analysis techniques. At the same time, the domain gap between web and unedited videos still limits vision-language applications to those between text and edited videos. Ego vision datasets are actively collected to overcome such problems; as another format of unedited videos, this paper provides a dataset with fixed-viewpoint unedited videos (FV videos). We can effortlessly obtain FV videos with commercial smartphones. We collected 145 videos, a total of 40 hours of footage, in which participants prepare foods based on given recipes. We manually add action graphs that tie videos and procedural texts while identifying the workflow of the process. In addition, we propose two benchmark tasks on this dataset: online recipe retrieval (OnRR) and dense video captioning on FV videos (DVC-FV). Experimental results demonstrated that recent SoTA methods can not solve OnRR and DVC-FV trivially.

1. はじめに

複数の材料を組み合わせて価値の高い製品を作り出すことは社会の物質的な豊かさを支える基本的な営みである。近年の機械学習技術の発達に伴い、このような活動の理解を促進するため、動画による観察から作業内容を理解するための研究が活発化している [1, 2, 3, 4]。現在の動画理解研究は HowTo100M[5] や YT-Temporal-1B [6] といった超大規模な Web 動画データセットによる事前学習に基づいている。しかし、Web 動画は人間に理解しやすいように注目すべき領域を大写しにしたり、あるいは、繰り返しのシーンを省略するように、時空間的に編集されている。このため、未編集の動画を対象にした作業内容理解との間には大きなドメインギャップが存在する。未編集動画の理解に関する研究は現在、主に一人称視点映像を対象に研究が進んでいる [2, 7]。本研究では、より広いユーザにアプローチしやすいスマートフォンによる観測を前提として、固定視点の未編集映像 (FV videos: Fixed-Viewpoint unedited videos) を対象とした新しいデータセットである COM Kitchens[*1] を構築する。このデータセットの特徴はスマートフォンをユーザに配布し、レシピに従った調理作業を収集することで、従来のデータセットにはない環境やユーザの多様性を実現している点にある。また、このデータセットを用いたベンチマーク課題として、オンラインレシピ検索（OnRR: Online Recipe Retrieval) という新しい問題、および、密な動画キャプショニング (DVC-FV: Dense Video Captioning for FV videos) という動画キャプショニングにおける新しい設定を提案する。

オンラインレシピ検索 (OnRR) は作業のある時点までの映像を対象とし、(1) その映像がどの指示文書に従っているかと (2) 指示文書のどの時点まで作業が進んでいるのかを同時に推定する課題である。OnRR を解くことができれば、作業内容に基づいたレシピの推薦や、作業に同期したナビゲーションといった課題を解くことができるようになる。一方、DVC-FV はオフラインの視覚言語問題として定義されている。OnRR が実用的な応用を意識しているのに対して、DVC-FV は Web 動画とのドメインギャップの解消に焦点を当てた課題として設計されている。

連絡先: 橋本敦史， atsushi.hashimoto@sinicx.com

*1 Cookpad OMRON Kitchens dataset の略称。

2. COM Kitchens データセット

データセットの構築にあたっては、一般家庭の調理台を簡易な三脚カメラに設置したスマートフォンで撮影する形式を取った。近年のスマートフォンカメラは広角化が進み、三脚程度の高さであっても、調理台を構成するコンロ、作業スペース、シンクの大半を撮影することが可能になってきている。さらに、スマートフォンのユーザインターフェイスは一般ユーザに十分に浸透しており、実験参加者が自ら設置や撮影操作、バッテリーの管理を行うことができる。このため、研究者が対面でのインストラクションや機器設置を行う必要がない。さらに、このように収集したデータセットに基づいて開発された技術は、そのままスマートフォンアプリとしての実装に耐えられる。

図 1 は実際に収集した動画のフレーム列と、それに対して Action Graph [1] をアノテーションした例である。Action Graph は映像につけた矩形、および、矩形間の関係を示す枝からなる。矩形は調理者が行う動作 (AP: Action by Person [14]) 毎に、その AP の前後の食材の状態に対してアノテーションを行った。これをそれぞれ動作前矩形 (Before)、動作後矩形 (After) と呼ぶ。さらに、複数の材料が混ぜられる場合を表現するため、混合が起きた場合のみ、動作目的地矩形 (Destination) を付与した。AP は手順書の動詞に相当するため、矩形と動詞の対応関係を ID を通して、映像と文書の関係を得ることができる。矩形間の枝は「材料名→動作前矩形」(I2B)、「動作前矩形→動作後矩形」(B2A)、「動作目的地矩形→動作後矩形」(D2A)、

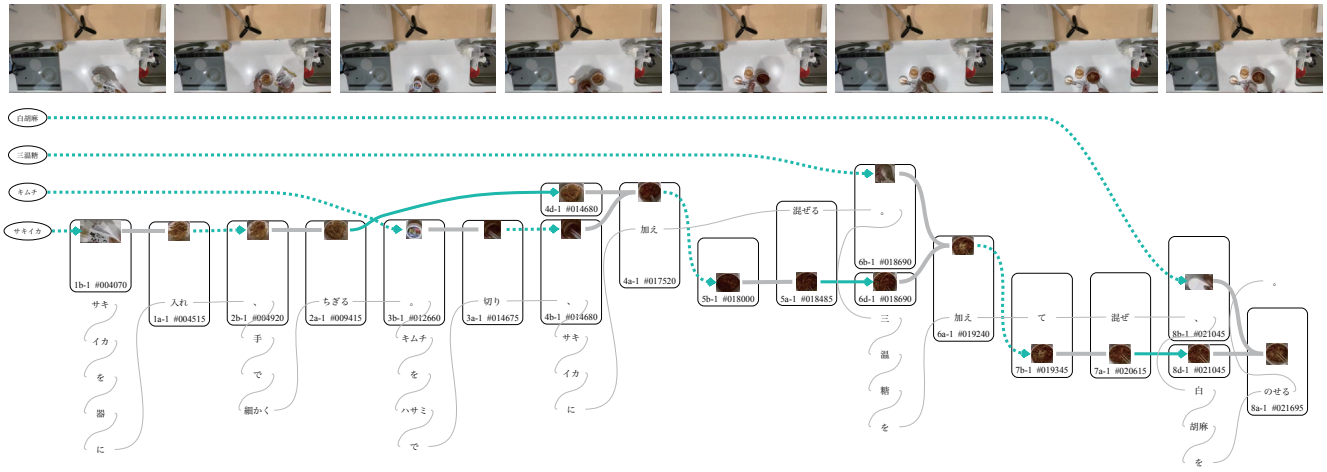

図 1: 映像 (上段) と付与された Action Graph の可視化 (下段) の例。Action Graph は作業者による動作 (AP) 毎に、動作前、動作後、および任意で動作目的地の 3 種類の矩形が付与される。それぞれの矩形は、その材料の同一性に基づいて枝によって紐づけられている。なお、AP は手順書の動詞と紐づいている。また、この可視化例はみやすさのために相当単純なものを選んでいる。

データセット	発表年	作業題材	#タスク	#環境	#動画	合計 (h)	平均 (m)	分節のタイプ	分節の記述方法
MMAC [8]	2008	Cooking	1	1	32	8	15.0	action	130 action cls.
MPII [9]	2012	Cooking	14	1	44	8	13.4	action	65 action cls.
50 salads [10]	2013	Cooking	2	1	50	5	5.4	action	51 action cls.
Breakfast [11]	2014	Cooking	-	18	1,712	77	2.7	action	10 action cls.
IKEA ASM [12]	2021	Furniture	4	5	371	35	5.7	action	noun+verb
Assembly101 [4]	2022	Assembly	15	1	1,425	167	7.1	act./step	1,380 action cls./noun+verb
EgoExo4D [13]	2023	Multiple	85	131	4,481	197	2.6	action	action cls. (fine/coarse)
COM Kitchens	Ours	Cooking	139	70	145	40	16.6	step	fine instructions

表 1: 手順書に従った作業を固定視点で観測した動画データセットの比較。 Assembly101 と EgoExo4D は公平のため、各 take につき 1 台のカメラのみを対象としたデータとなっている。

「動作後矩形→動作前矩形」(A2B) の 4 種類を用意した。調理では材料が混ぜられるため、一般的な物体検出のようなカテゴリを定義することが困難である。I2B 枝は、従来の物体カテゴリの情報をグラフに与える。B2A および D2A 枝は AP に対応する各動作に対して一意に決まる枝で、動作の前後の物体の状態変化を表現するとともに、動作区間のアノテーションにもなっている。A2B はある AP の動作前矩形が、直前にどの AP で使われたものかの対応関係を与える。これにより、材料が手順書の工程を経て完成品にいたるまでを追跡することができる。なお、動画毎に、作業者が従った手順書が存在するが、手順書を完全に遵守できた例が稀であるため、それらの手順書とは別に、アノテーターが動画を元に手順書を書き起こし、アノテーションにはその書き起こしレシピを用いた。

以上のデータ収集、および、アノテーションを経て、70 人の実験参加者から、139 種類のレシピを対象とした 145 動画、合計 40 時間 (平均 16.6 分) の映像を得た (表 1)。なお、全ての実験参加者が異なる環境で調理を行っている。また、アノテーションの結果、全フレームの 80%が AP の動作中となり、特に複数の AP が重複するものは 15%であった。指示文書中の AP の数は合計 2,286 であり、それが動作中で実行された回数は合計 2,852 区間、平均継続時間は 46.7 秒となった。また、矩形の総数は 6826 個、枝の総数は 8061 本となった。これは 1 動画あたり、平均で 47.1 個の矩形、55.6 本の枝が付与されていることに相当する。

表 1 に既存のデータセットとの比較を載せる。COM Kitchens データセットの特徴は、タスク数と環境数が多いこと、および、作業時間が長いことにある。タスク数は動画中に行われた作業に対応する手順書の数を表している。また環境数は撮影環境を表しており、COM Kitchens においては家庭のキッチンの数がそれにあたる。なお、同データセットでは各キッチンにちょうど一人の調理者が対応しているため、環境数と作業者数は同一である。EgoExo4D は多くの研究室が合同で収集したデータセットであるが、COM Kitchens が小規模なチームで同等の多様性を実現できたのはスマートフォンを配布し、対面でのインストラクションを経ずにデータを収集するという方法によるところが大きい。

3. ベンチマーク課題

以下では COM Kitchens とともに提案する 2 つのベンチマーク課題について、その設定とベースラインとなる実験結果を紹介する。いずれの実験においても 145 の動画を 90/26/29 の訓練/検証/テストセットへ分割して用いた。分割に際しては観測環境が漏洩しないように配慮した。

3.1 オンラインレシピ検索 (OnRR)

OnRR は以下の 2 つの部分課題からなる。

i) 実現可能レシピ検索課題 (Feasible Recipe Retrieval)

Task	Method	Metrics			
		R@1	R@5	R@10	MdR
Feasible Recipe Retrieval	Random	1.8	8.6	15.8	-
	UniVL [16]	3.4	5.7	9.2	227.0
	CLIP4Clip [17]	0.0	0.0	10.3	79.0
	X-CLIP [18]	0.0	6.8	10.3	89.0
Recipe Stage Retrieval	Random	6.3	31.6	63.3	8.0
	UniVL [16]	17.2	48.2	68.9	5.0
	CLIP4Clip [17]	6.8	48.2	68.9	5.0
	X-CLIP [18]	10.3	51.7	68.9	4.0

表 2: Comparison between online recipe retrieval (OnRR) performances of baseline models. R@K and MdR represent recall at rank K (↑) and median rank (↓), respectively. This table provides only the early-stage setting (using the first 25% of the video as input); results in other settings are detailed in the supplementary.

ii) レシピ段階検索課題 (Recipe Stage Retrieval)

部分課題 (i) における実現可能レシピ (Feasible Recipe) とは、映像のある時点を対象として、その段階からレシピに従った調理へ移行すると仮定した場合に、手戻りなく移行が可能なレシピのことである。例えば、ユーザが特に具体的なレシピをシステムに申告することなく作業を開始した後で、システムから作業内容に基づいてレシピを推定または推薦するような応用を想定している。

部分課題 (ii) はレシピが既知である、または部分課題 (i) が解けている場合を想定し、さらに映像のある時点でレシピ中のどの段階まで実行済みかを特定する。これは、進捗に応じた情報提示を行う応用を想定している。

いずれの部分課題に対しても、各映像を開始時刻から 25%,50%,75%,100%時点までの 4 つの部分映像 (段階 1〜4) に分割し、段階毎に評価を行う。なお、部分課題 (i) は、収集した映像データをクエリとして、レシピを検索する。しかし、単に COM Kitchens のテストセットのレシピのみを検索対象のレシピプールとする場合、各映像に対応する実現可能レシピの数が常にほぼ 1 となってしまう。このため、Cookpad Recipe Dataset(CRD)[15] を用いて、検索対象レシピプールを拡張した。具体的には、テストセットのレシピの材料に基づいて、内容が類似する 1,828 レシピを選出した。これにより、各段階ごとに 991/243/19/5 の実現可能レシピを新たに得た。

表 2 にこの課題に対するベースラインの結果を示す。紙幅の都合から、ここでは 25%時点までの部分映像に対する結果のみを示す。部分課題 (i) については、従来手法はテキスト全体を考慮してしまうためか、ほとんど検索課題を解くことができなかった。一方、部分課題 (ii) については一定の精度が達成できたものの、やはり実用に足る精度を達成することはできなかった。これら 2 つの精度を向上することが直接的に作業支援システムの実現に繋がると考えており、今後、これらのベンチマークに対してデータやアルゴリズムの両面からの発展を期待している。

3.2 固定視点未編集動画キャプショニング (DVC-FV)

DVC-FV は技術としては単に DVC を未編集固定視点映像に適用するだけのものである。この課題の目的は Web 動画で学習したモデルが、どの程度、ドメインギャップの影響を受けるのかを知ることであり、また、それを将来の技術向上につなげることにある。特に、未編集映像と編集済み映像の違いは映像中の重要領域が最初からハイライトされているのか、モデル自身がハイライトすべき点を見つけられるかにある。このため、Transformer の注意 (Attention) に着目した解析を行った。

Model	FT	AG	S	C	M
Vid2Seq [19]	-	-	0.017	0.066	0.010
Vid2Seq	✓	-	0.369	2.832	0.642
Vid2Seq	✓	RL	0.211	1.381	0.285
Vid2Seq	✓	AS	0.266	2.513	0.423
Vid2Seq	✓	RL+AS	0.581	6.195	1.142

表 3: DVC-FV の性能比較。S/C/M はそれぞれ SODA_c(↑)、CIDEr(↑) 、 および METEOR(↑) の評価値を百分率 (%) 形式で表している。また、FT の列は追加学習の有無を、AG の列は Action Graph の利用方法をそれぞれ表している。

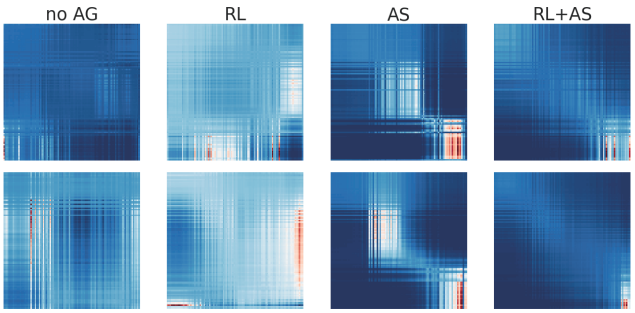

図 2: エンコーダの最初と最後の層の Attention の可視化。対象のサンプルはランダムに選出した。

まず従来手法として、大量の Web 動画に基づいた事前学習を行うことで Web 動画の DVC に対して SoTA となっている Vid2Seq[19] を用いた。また、注意先を積極的に改善する方法として、Action Graph を Relation Label (RL)[20] と Attention Supervision (AS) として用いる手法 [21]、および、それら 2 つを同時に用いる手法を作成し、比較した。

表 3 にその結果を示す。全体的にスコアが非常に低く、Web 動画と未編集固定視点映像のドメインギャップが非常に大きいことが示されている。その中にあって、追加学習が一定の効果を持っていることが確認できる。Action Graph の利用方法に関して、RL, AS いずれも単体では単純な追加学習よりも精度が下がる結果となった。一方で、両方を併用することで単純な追加学習に勝る精度を得ることができた。

この現象を理解するため、各ベースラインに対して注意を可視化した (図 2)。RL と AS はいずれも、注意を特定のフレームに集中させる傾向が見られるのに対して、併用した場合にのみ、時間的に近いフレームに適度に注意が分散していることがわかる。このような注意の適度な平滑化効果が生まれた結果、適度に文脈を考慮したキャプションができるようになったものと考えられる。

4. おわりに

本研究では、スマートフォンで撮影した新しい言語リソース付きの動画データセットである COM Kitchens を提案した。また、Action Graph アノテーションを活用した 2 つの課題を

提案し、そのベースラインを示した。いずれの課題においてもWeb 動画に強く依存したモデルでは高い精度を出すには至らないことを確認した。また、これらの結果から、スマートフォンで動く実用的なアプリケーションを創出するためには、Web 動画と本データセットのドメインギャップを埋めるような技術開発が必要であるという知見を得た。

参考文献

[1] De-An Huang, Joseph J. Lim, Li Fei-Fei, and Juan Carlos Niebles. Unsupervised visual-linguistic reference resolution in instructional videos. In *CVPR*, pages 2183–2192, 2017.

[2] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The EPIC-KITCHENS dataset. In *ECCV*, 2018.

[3] Taichi Nishimura, Kojiro Sakoda, Atsushi Hashimoto, Yoshitaka Ushiku, Natsuko Tanaka, Fumihito Ono, Hirotaka Kameko, and Shinsuke Mori. Egocentric biochemical video-and-language dataset. In *ICCV Workshop*, pages 3129–3133, October 2021.

[4] F. Sener, D. Chatterjee, D. Shelepov, K. He, D. Singhania, R. Wang, and A. Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *CVPR*, pages 21064–21074.

[5] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019.

[6] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. MERLOT Reserve: Neural script knowledge through vision and language and sound. In *CVPR*, pages 16375–16387, June 2022.

[7] Kristen Grauman, Andrew Westbury, Eugene Byrne, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*, pages 18995–19012, June 2022.

[8] Ekaterina H Spriggs, Fernando De La Torre, and Martial Hebert. Temporal segmentation and activity classification from first-person sensing. In *CVPRW*, pages 17–24, 2009.

[9] Bernt Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, page 1194–1201, 2012.

[10] Sebastian Stein and Stephen J. McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *ACM international joint conference on Pervasive and ubiquitous computing*, page 729–738, 2013.

[11] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *CVPR*, pages 780–787, 2014.

[12] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The IKEA ASM dataset: Understanding people assembling furniture through actions, objects and pose. pages 847–859, 2021.

[13] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, et al. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives, 2023.

[14] Shinsuke Mori, Hirokuni Maeta, Yoko Yamakata, and Tetsuro Sasada. Flow graph corpus from recipe texts. pages 2370–2377, 2014.

[15] Jun Harashima, Michiaki Ariga, Kenta Murata, and Masayuki Ioki. A large-scale recipe and meal data collection as infrastructure for food research. pages 2455–2459, 2016.

[16] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. UniVL: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.

[17] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval. *Neurocomputing*, 508:293–304, 2022.

[18] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-CLIP: End-to-end multi-grained contrastive learning for video-text retrieval. In *ACM MM*, pages 638–647, 2022.

[19] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2Seq: Large-scale pretraining of a visual language model for dense video captioning. In *CVPR*, 2023.

[20] Youmi Ma, Tatsuya Hiraoka, and Naoaki Okazaki. Joint entity and relation extraction based on table labeling using convolutional neural networks. In *Workshop on Structured Prediction for NLP*, pages 11–21, 2022.

[21] Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. Jointly learning to align and translate with transformer models. In *and*, pages 4453–4462, Hong Kong, China, 2019.