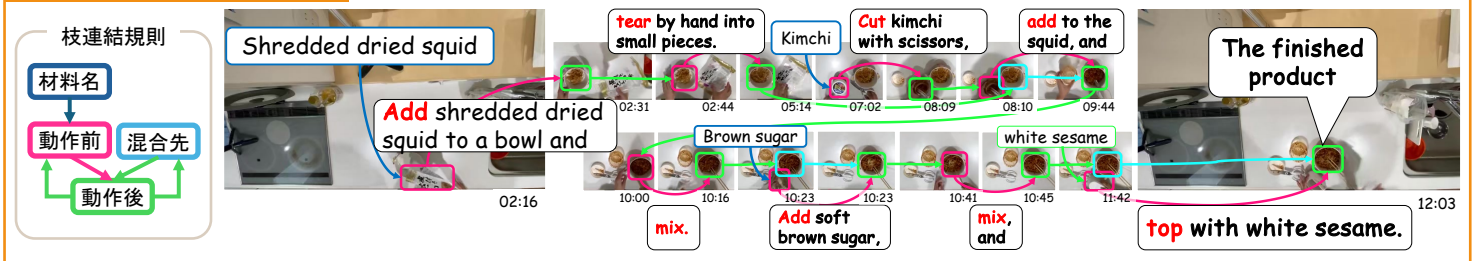


COM Kitchens: 調理作業理解のための言語資源つき固定視点映像データセットの構築

目的: Web動画ではなく「未編集なユーザ撮影映像」を対象とする作業理解研究基盤を作成

- 概要: 1. 一般家庭のスマホ撮影映像を収集 (70環境/139レシピ/145動画/平均16.6分)
2. 「Visual Action Graph」 (下図) で手順文章と映像を紐づけ

Visual Action Graph



調理者が各家庭で設置して撮影

iPhone 11 Pro+三脚を実験参加者に郵送
→研究者の立ち会いなしでデータ収集可能



カメラ設置例

画角は調理台ほぼ全域

Word Clouds (動詞・材料名)



レシピ・環境の多様性が高い+長時間動画

dataset	year	topic	tasks	# env.	# videos	total (h)	avg. (m)	seg. type	seg. description
MMAC [45]	2008	Cooking	1	1	32	8	15.0	action	130 actions cls.
MPII [39]	2012	Cooking	14	1	44	8	13.4	action	65 actions cls.
ACE [43]	2012	Cooking	5	1	35	2	3.6	action	8 actions cls.
50 salads [46]	2013	Cooking	2	1	50	5	5.4	action	51 actions cls.
Breakfast [22]	2014	Cooking	-	18	1,712	77	2.7	action	10 actions cls.
IKEA ASM [3]	2021	Furniture	4	5	371	35	5.7	action	noun+verb (n+v)
Assembly101 [40]	2022	Assembly	15	1	4,321	513	7.1	act./step	1,380 act. cls./n+v
COM Kitchens Ours		Cooking	<u>139</u>	<u>70</u>	145	40	<u>16.6</u>	act./step	instructional text

表1. 従来の固定視点映像データセットとの比較

数少ない「固定視点映像+言語資源」データセット

dataset	year	type	topic	tasks	# videos	total (h)	avg. (m)	seg. description	interval	# seg.
YouCookII [56]	2018	Web Cook.		89	2,000	176	5.3	coarse instruction	manual	4,325
ProeL [10]	2019	Web Multi.		12	720	47	3.9	coarse instruction	manual	498
COIN [47]	2019	Web Multi.		180	11,827	476	2.4	coarse instruction	manual	46,354
CrossTask [57]	2019	Web Multi.		83	4,700	376	4.8	coarse instruction	manual	19,278
MMAC-Captions [33]	2021	Ego Cook.		5	146	16	13.4	coarse instruction	manual	5,002
Epic Kitchens [7]	2022	Ego Cook.		70	700	100	8.6	narration	utterance	39,596
Ego4D [14]	2022	Ego Open		-	-	3,670	-	narration	manual*	-
BioVL2 [38]	2022	Ego Bio.		5	32	3	5.3	fine instruction	manual	408
VRF [44]	2022	Web Cook.		200	200	2	0.7	fine instruction	manual	3,705
COM Kitchens Ours		<u>FV</u> Cook.		<u>139</u>	<u>145</u>	<u>40</u>	<u>16.6</u>	fine instruction	manual	<u>2,852</u>

表2. 言語資源つき動画データとの比較

※厳密にはEgoExo4Dもあるが統計情報が集計されていないため表からは除外

ベンチマーク課題① オンラインレシピ検索 (新提案課題)

入力: 調理途中までの動画

出力: 調理中のレシピ + 進捗状況の特定

Feasible Recipe Retrieval Recipe Stage Identification

Task	Method	Early (25%)				Middle (50%)			
		R@1	R@5	R@10	MdR	R@1	R@5	R@10	MdR
Feasible Recipe Retrieval	Random	1.8	8.6	15.8	-	0.4	1.8	3.1	-
	UniVL [25]	3.4	5.7	9.2	227.0	3.4	5.7	9.2	231.0
	CLIP4Clip [26]	0.0	0.0	10.3	79.0	0.0	0.0	6.8	85.0
	X-CLIP [27]	0.0	6.8	10.3	89.0	0.0	3.4	3.4	320.0
Recipe Stage Identification	Random	6.3	31.6	63.3	8.0	6.3	31.6	63.3	8.0
	UniVL [25]	17.2	48.2	68.9	5.0	9.2	63.3	89.2	3.0
	CLIP4Clip [26]	6.8	48.2	68.9	5.0	10.3	55.1	86.2	4.0
	X-CLIP [27]	10.3	51.7	68.9	4.0	17.2	37.9	93.1	6.0

ベンチマーク課題② Dense Video Captioning

Visual Action Graphを教師信号として利用

RL: Relation Label を同時推定

AS: TransformerのAttention先を教師あり学習

Model	FT	AG	SODA_c(↑)	CIDEr(↑)	METEOR(↑)
PDVC [49]	-	-	0.022	0.000	0.000
Vid2Seq [52]	-	-	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

橋本 敦史 / オムロンサイニックス株式会社 / PI

* 本データセットはクックパッド株式会社との共同研究により収集したものをオムロンサイニックス社が研究用途に限ってIDRを通して公開するものです。
* 本研究はJST ムーンショット型研究開発事業 JPMJMS2236 および JSPS科研費 21H04910 の助成を受けたものです。