EgoOops!データセット：手順書に従う作業の一人称視点映像への作業誤りアノテーション

羽路 悠斗 [1]　西村 太一 [2]　山本 航輝 [1]　梶村 恵矢 [1]　崔 泰毓 [1]
亀甲 博貴 [3]　森 信介 [3]
[1] 京都大学大学院 情報学研究科　[2] LINE ヤフー株式会社
[3] 京都大学 学術情報メディアセンター
[1]{haneji.yuto.58c,yamamoto.koki.76n}@st.kyoto-u.ac.jp
[1]{kajimura.keiya.48x,cui.taiyu.33c}@st.kyoto-u.ac.jp
[2]tainishi@lycorp.co.jp　[3]{kameko,forest}@i.kyoto-u.ac.jp

概要

本研究の目的は，手順書に従う作業の一人称視点映像を，手順書を参照しながら解析し，作業の誤りの時間区間と種類を予測することである．まず，手順書に従う 5 つの作業，合計 50 本，6.8 時間の一人称視点の作業映像を撮影し，EgoOops!データセットを構築する．映像には，事前に定義した誤りの種類と，詳細な説明文を付与している．また，大規模な一人称視点映像と言語のペアで事前学習したモデルを用いて，手順認識と作業誤り分類を統合的に行う手法を提案する．手順認識で 36.9，作業誤り分類で 22.3 の F1 スコアであった．映像のみを用いた結果との比較から，手順書の参照が作業誤り分類に有効だとわかった．

1　はじめに

料理や組み立てなどの手順書に従う作業において，動作の誤りや物体の取り違えなどの作業の誤りが起こる．そうした誤りにより，作業の質が低下したり危険が増したりする．対策としては，よく起こる誤りを特定して作業者を教育したり，誤りの発生を警告したりすることが挙げられる．こうした対策には多くの人員を要するので，作業の様子を撮影して自動で誤りを検出することが望ましい．

既存研究では，映像からの作業誤りの検出を目的として，撮影した作業映像の誤りをアノテーションしたデータセットが構築されてきた [1, 2, 3, 4, 5]．これらの映像からの誤り検出には，動作認識 [1, 2, 3, 6] や映像異常検知 [4] の手法が用いられてきた．ここで，従来のデータセットの撮影では，手順を厳密には定めないことが多かったため [1, 2, 3, 5]，手順書を参照するモデルは実現されていない．

本研究の目的は，一人称視点の作業映像を手順書も参照しながら解析し，手順から逸脱する作業の誤りの時間区間と種類を予測することである．まず，多様な領域に渡る 5 つの作業を手順書に従って実施し，合計 50 本，6.8 時間の一人称視点の作業映像を収集し，EgoOops!データセットを構築する（図 1）．手順区間，手順ラベル，事前に定義した 6 種類の作業誤りのラベル，誤りの詳細な説明文を映像に付与しており，作業誤りを検出するモデルの学習と評価に利用できる．また，大規模な一人称視点映像と言語のペアで事前学習したエンコーダによって，作業映像と手順書を埋め込み，手順の認識と作業誤りの分類を統合的に行う手法を提案する．

EgoOops!データセットを用いて評価実験を行う．手順区間の特定において，動作区間検出の最高性能のモデルの mAP は 52.24 であった．提案手法による手順認識の F1 スコアが 36.9，作業誤り分類の F1 スコアが 22.3 であった．作業誤りの分類で映像のみを用いると F1 スコアが大きく低下したため，手順書の参照が有効だといえる．

2　EgoOops!データセットの構築

本研究では，手順書に従う作業の誤り検出のために，EgoOops!データセットを構築する．本データセットは，一人称視点映像と手順書からなる．

2.1　手順書の用意

映像の収録前に，基準に沿って作業を選定し，手順書を用意する．

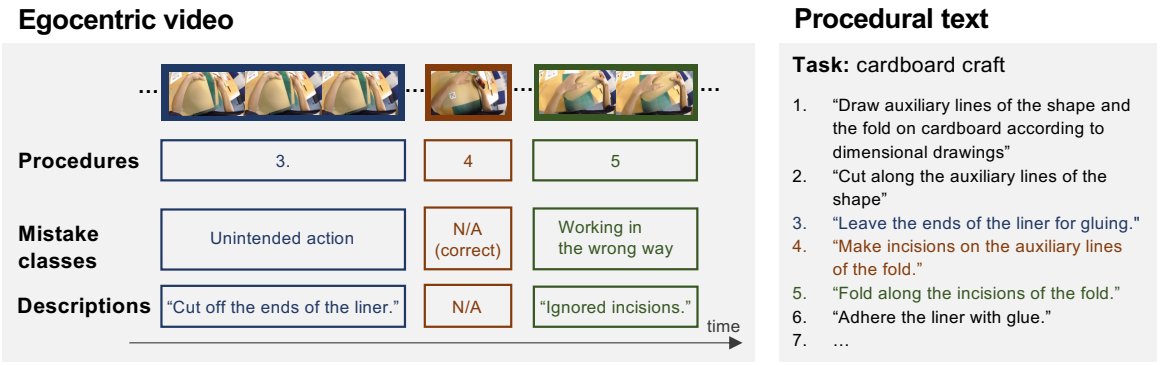


図 1: EgoOops!データセットの例．手順書に従う作業の一人称視点映像に，手順区間，手順ラベル，事前に定義した 6 種類の作業誤りのラベル，誤りの詳細な説明文を付与する．

図 2: 作業の撮影風景．

作業の選定　Web を検索して手順書に従う作業を収集し，電気回路，光の混色実験，イオン反応実験，おもちゃの積み木作り，ダンボール工作の 5 つの作業を選択した．作業の領域，誤りの種類，映像の長さ，物体，動作の多様性が，選定の基準である．

手順書の用意　選定した作業について，手順書を用意する．Web 上や専用のキットに付属の手順書がある作業では，利用する道具を明記するなど，著者のうち 1 名がより詳細に書き換える．既存の手順書がない作業では，著者のうち 1 名が事前に作業して手順を書き起こしておく．日本語で手順書を用意し，英文で事前学習したモデルを適用できるように，手作業で英語版の手順書に翻訳もする．用意した手順書を参照して，収録参加者は作業を行う．

2.2　映像の収録

収録の工程　男性 4 名の日本人の大学院生の参加者に，手順書に従って作業してもらう．参加者にそれぞれの作業を 2 本か 4 本ずつ割り当て，各作業を 10 本ずつ撮影する．多くの作業誤りを撮影するために，我々が事前に用意したか参加者が考えた誤りを 5 本の映像では意図的に含め，他の 5 本の映像では手順書通りに作業する．

撮影環境　撮影風景を図 2 に示す．Panasonic HX-A500 一人称視点カメラを，参加者の頭部に装着して撮影する．道具や手順書など作業に用いる物を，予め机に置いておく．作業の様子を近くで撮影し，画角の変動を抑えるために，座って作業する．

2.3　アノテーションの規格

収録した映像に，手順区間の開始時刻と終了時刻を，手順のラベルと，作業誤りの種類のラベルと説明文とともに付与する（図 1）．手順のラベルは順序関係の誤りに，作業誤りの種類のラベルは手順の実行内容の誤りにそれぞれ対応する．本稿の実験では，作業誤りの分類に取り組む．

区間　区間の開始時刻はいずれかの物体を掴んで手順を始める時点，終了時刻は手順を終えて全ての物体を離す時点とする．ただし，ある物体を持ったまま次の手順に移行する場合には，それ以外の物体を掴んだり離したりする時点とする．なお，ここでいう手順には，手順書で指示されていないが作業者が誤って実行したものもある．

手順　手順書の手順には実行順のラベルが付いており，それらを参照して各区間の手順のラベルを付与する（図 1）．ただし，手順書にない手順は，**未定義** とする．なお，手順の重複，欠落，順番の入れ替わりが起こることがある．こうした順序関係の誤りの検出は，本稿では扱わず今後の課題とする．

作業誤りの種類と説明文　作業誤りを含む区間には，事前に定義した誤りの種類のラベルを付与する．さらに，どう誤っているかを具体的に説明するために，詳細な内容を英文で記述する．定義した誤りの種類は，1. 誤った物体で手順を実行する，2. 誤った物体を掴み，使わず手放す，3. 作業誤りの訂正，4. 意図しない行動，5. 誤った方法で手順を実行する，6. その他の 6 種類である．なお，複数の種類に該当する場合も，その他に分類する．本稿の実験では，誤りの種類の予測に取り組み，将来的には

表 1: 手順数と手順あたりの単語数を作業間で比較.

作業	手順数	手順あたりの単語数
電気回路	8.0	7.6
光の混色	8.0	17.0
イオン反応	9.0	12.7
積み木	7.0	18.6
工作	14.0	9.6

表 2: 映像の平均長，映像あたりの区間数，区間の平均長を作業間で比較．各作業 10 本ずつ撮影した．

作業	映像	区間	
	平均分数	映像あたりの区間数	平均秒数
電気回路	3.2	9.6	15.6
光の混色	4.4	8.8	26.7
イオン反応	5.4	9.5	29.7
積み木	1.9	8.7	8.9
工作	26.1	16.5	87.7
合計	8.2	10.6	41.3

説明文を応用（例：自動生成）することを目指す．

2.4 統計情報

EgoOops!データセットの統計情報を，手順書・映像・作業誤りについて算出する．また，別のアノテータによるアノテーションとの一致率を評価し，アノテーションの質を確かめる．

手順書 表 1 に示すように，手順数や手順ごとの単語数は作業ごとに異なり，手順書の複雑さは様々だといえる．手順数は，積み木が最も少なく（7 手順），ダンボールが最も多い（14 手順）．手順ごとの単語数は，電気回路が最も少なく（平均 7.6 単語），積み木が最も多い（平均 18.6 単語）．

映像 表 2 に示すように，映像長・区間長・区間数の観点で多様な映像を含んでいる．積み木がいずれも最小（平均 1.9 分の映像長，平均 8.9 秒の区間長，1 映像あたり平均 8．7 区間）であり，ダンボールがいずれも最大（平均 26.1 分の映像長，平均 87.7 秒の区間長，1 映像あたり平均 16.5 区間）である．

表 3: 各種類の誤りの数を作業間で比較. 種類の定義は 2.3 節を参照.

作業	誤りの種類					
	1.	2.	3.	4.	5.	6.
電気回路	2	5	0	1	2	2
光の混色	2	5	0	0	5	2
イオン反応	4	3	1	2	1	9
積み木	2	4	5	1	4	0
工作	2	2	0	0	2	4

表 4: 区間の特定の結果.

手法	mAP					
	0.1	0.2	0.3	0.4	0.5	avg
Shan *et al.* [8]	4.31	2.86	1.12	0.56	0.22	1.81
ActionFormer [9]	55.56	48.20	31.21	17.35	7.42	31.95
TriDet [10]	83.24	71.18	54.53	35.01	17.24	*52.24*

作業誤り 表 3 に示すように，多様な種類の誤りを含んでいる．作業ごとに特有の誤りの傾向（例：イオン反応では，誤った物体で手順を実行することが多い）があり，その傾向は作業間で多様である．

アノテーションの一致率 アノテーションの質を確かめるために，別の 1 人のアノテータが新たに付与するアノテーションとの一致率を評価する．アノテータが付与した区間と手順のラベルとの一致率は，共通部分を和集合で割った tIoU で 84.2 である．それとは別に，EgoOops!データセットのアノテーションの区間に，アノテータが作業誤りの種類のラベルと説明文を付与する．種類の一致率は F1 スコアで 83.3 であり，説明文の一致率は BERTScore [7] で 96.3 である．これらの一致率から，アノテーションが高品質だといえる．

3 実験

2 節で構築した EgoOops!データセットを用いて，作業誤り検出の評価実験を行う．エンドツーエンド認識にはデータ量が少ないため，1. 区間の特定，2. 手順の認識，3. 作業誤りの分類に分ける．

3.1 区間の特定

問題設定 映像全体の一連の T 枚のフレーム $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$ が与えられ，N 個の区間 $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N\}$ を出力する．区間 $\mathbf{y}_i = (s_i, e_i)(i \in [1, N])$ は，開始時刻 s_i と終了時刻 e_i からなり，$s_i \in [1, T]$, $e_i \in [1, T]$, $s_i \leq e_i$ である．動作区間検出で広く使われている mAP で評価する [11, 12, 13, 14]．共通部分を和集合で割った tIoU が閾値以上のとき，予測区間が正解区間と一致しているとみなす．

手法 動作区間検出において最高性能である，ActionFormer [9] と TriDet [10] を評価する．また，手順の実行中には物体を操作するため，hand-object detector [8] で手と物体の接触を検出したフレームが連続する区間を評価する．

結果 tIoU の閾値 $\{0.1, 0.2, 0.3, 0.4, 0.5\}$ における評価結果と，それらの平均を表 4 に示す．TriDet に

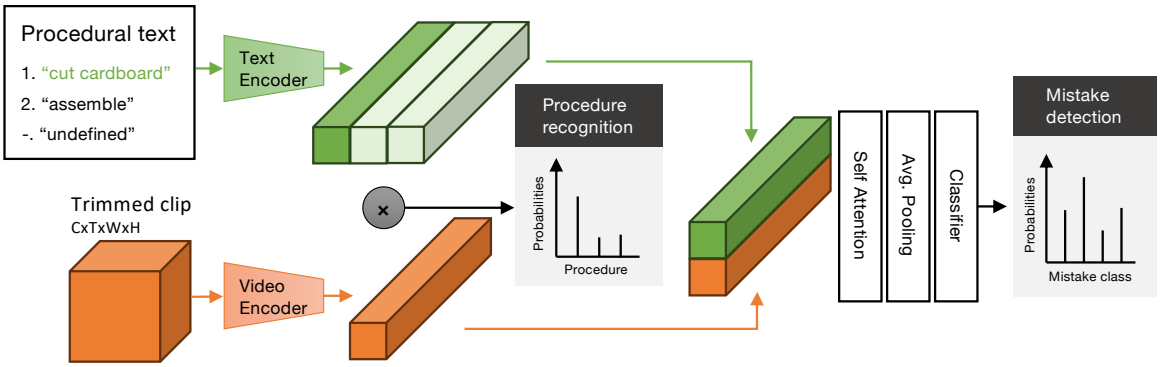


図 3: 提案手法の概要.

表 5: 手順の認識の結果．タスクごとにマクロ平均を算出し，さらにそれらの平均を示す．

エンコーダ	Precision	Recall	F1
EgoVLP [15]	39.6	40.0	36.9

表 6: 作業誤りの分類の結果．マクロ平均を示す．

エンコーダ	モダリティ	Precision	Recall	F1
EgoVLP [15]	映像	6.7	32.7	10.3
EgoVLP [15]	映像と手順書	19.1	35.1	22.3

よる予測の mAP の平均は 52.24 であり，最も高精度である．EPIC-KITCHENS-100 データセット [13] の動作区間検出において，TriDet のスコアは 25.4 であり [10]，EgoOops!データセットの区間は比較的検出しやすいといえる．

3.2 手順の認識と作業誤りの分類

タスク定義 アノテーションに従い時間方向にトリミングした M_i 枚のフレームからなる区間 $\mathbf{V}_i = \{\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_{M_i}\}$，$N$ 手順の手順書 $\mathbf{T} = \{t_1, t_2, ..., t_N\}$，$C$ 個の誤りの種類 $\{m_1, m_2, ..., m_C\}$ が与えられる．t_{N+1} = **未定義** を含めた $N+1$ 手順の一つ $t_k (k \in [1, N+1])$ と，m_{C+1} = **正解** を含めた $C+1$ 種類の一つ $m_l (l \in [1, C+1])$ への分類を出力する．A 個の区間が一つの映像 $\mathbf{V} = \{\mathbf{V}_1, \mathbf{V}_2, ..., \mathbf{V}_A\}$ には含まれる．Precision, Recall, F1 スコアで評価する．

手法 図 3 に示すように，映像と言語の埋め込みのコサイン類似度で手順を認識し，埋め込み空間上の分類層により作業誤りを分類することを提案する．埋め込みには，大規模な一人称視点映像と言語のペアからなるデータセット Ego4D [14] で事前学習した EgoVLP [15] を用いる．本稿では，手順認識と誤り分類を分けて評価するために，アノテーションされた手順の文の埋め込みを誤りの分類に用いる．

区間 $\mathbf{V}_i$ のフレーム系列を，ベクトル $\hat{\mathbf{V}}_i \in \mathbb{R}^d$ に埋め込む．各手順 $t_j (j \in [1, N+1])$ の文を，ベクトル $\hat{\mathbf{t}}_j \in \mathbb{R}^d$ にそれぞれ埋め込んで，手順埋め込み系列 $\hat{\mathbf{T}} = \{\hat{\mathbf{t}}_1, \hat{\mathbf{t}}_2, ..., \hat{\mathbf{t}}_{N+1}\}$ を得る．映像の埋め込みとのコサイン類似度が最も大きい手順を認識する．

区間 $\mathbf{V}_i$ の映像の埋め込み $\hat{\mathbf{V}}_i$ とアノテーションされた手順 $t_{gt_i} (gt_i \in [1, N+1])$ の文の埋め込み $\hat{\mathbf{t}}_{gt_i}$ を積み重ねて，$\mathbf{F}_i \in \mathbb{R}^{2 \times d}$ を得る．self-attention 層と続く average pooling 層により一次元目の方向に融合して，$\mathbf{f_i} \in \mathbb{R}^d$ を得る．任意の分類器により，作業誤りの種類 m_l に分類する．

設定 EgoVLP の重みも学習する．埋め込みの次元 $d = 256$ とする．分類器は 1 層の線形層とする．

結果 表 5 に，手順の認識の結果を示す．Precision は 39.6, Recall は 40.0, F1 スコアは 36.9 である．

表 6 に，作業誤りの分類の結果を示す．映像と手順書を用いると，Precision は 19.1, Recall は 35.1, F1 スコアは 22.3 だった．F1 スコアの 7 クラス分類におけるチャンスレートの約 14.3 より高性能である．また，映像のみを用いると F1 スコアが 12.0 低下したため，手順書の参照が有効だといえる．

4 終わりに

本研究では，手順書に従う作業の誤り検出のために，一人称視点映像と手順書からなる EgoOops!データセットを構築した．手順書に従う作業を対象とする点，誤りの種類ラベルも付与している点，作業の領域と誤りの種類が多様な点が貢献である．

収集した作業映像に対して，大規模な一人称視点映像と言語のペアで事前学習したエンコーダを用いて，手順認識と作業誤り分類を統合的に行った．作業誤りの分類では，提案手法の性能がチャンスレートを上回った．また，映像のみを用いた結果との比較により，手順書の参照が有効だとわかった．

参考文献

[1] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A Large-Scale Multi-View Video Dataset for Understanding Procedural Activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21064–21074, June 2022.

[2] Reza Ghoddoosian, Isht Dwivedi, Nakul Agarwal, and Behzad Dariush. Weakly-supervised action segmentation and unseen error detection in anomalous instructional videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10128–10138, 2023.

[3] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. HoloAssist: An egocentric human interaction dataset for interactive AI assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20270–20281, 2023.

[4] Rohith Peddi, Shivvrat Arya, Bhrath Challa, Likhitha Pallapothula, Akshay Vyas, Qifan Zhang, Jikai Wang, Vasundhara Komaragiri, Eric Ragan, Nicholas Ruozzi, Yu Xiang, and Vibhav Gogate. Put on your detective hat: What's wrong in this video? In *Proceedings of the 40th International Conference on Machine Learning*, Honolulu, Hawai, USA, 2023.

[5] Tim J. Schoonbeek, Tim Houben, Hans Onvlee, Peter H. N. de With, and Fons van der Sommen. IndustReal: A Dataset for Procedure Step Recognition Handling Execution Errors in Egocentric Videos in an Industrial-Like Setting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4365–4374, 2024.

[6] Guodong Ding, Fadime Sener, Shugao Ma, and Angela Yao. Every Mistake Counts in Assembly, July 2023.

[7] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations*, April 2020.

[8] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F. Fouhey. Understanding Human Hands in Contact at Internet Scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9866–9875, Seattle, WA, USA, June 2020. IEEE.

[9] Chen-Lin Zhang, Jianxin Wu, and Yin Li. ActionFormer: Localizing Moments of Actions with Transformers. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Proceedings of the European Conference on Computer Vision*, Vol. 13664, pp. 492–510, Cham, 2022. Springer Nature Switzerland.

[10] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. TriDet: Temporal Action Detection With Relative Boundary Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18857–18866, 2023.

[11] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 961–970, 2015.

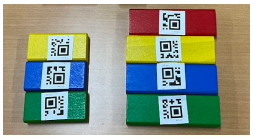
[12] Haroon Idrees, Amir R. Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The THUMOS challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, Vol. 155, pp. 1–23, February 2017.

[13] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling Egocentric Vision: Collection, Pipeline and Challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision*, Vol. 130, No. 1, pp. 33–55, January 2022.

[14] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18973–18990, New Orleans, LA, USA, June 2022. IEEE.

[15] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z. XU, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, Chengfei Cai, WANG HongFa, Dima Damen, Bernard Ghanem, Wei Liu, and Mike Zheng Shou. Egocentric video-language pretraining. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 35, pp. 7575–7586. Curran Associates, Inc., 2022.

(a) 銅（左），亜鉛（中央），マグネシウム（右）の外観の違い．亜鉛とマグネシウムを見た目から判別するのは困難である．



(b) マイクロ QR コードを添付した様子．短い直方体（左）と長い直方体（右）は類似している．

図 4: 判別が困難な物体があるため，マイクロ QR コードを物体に添付．

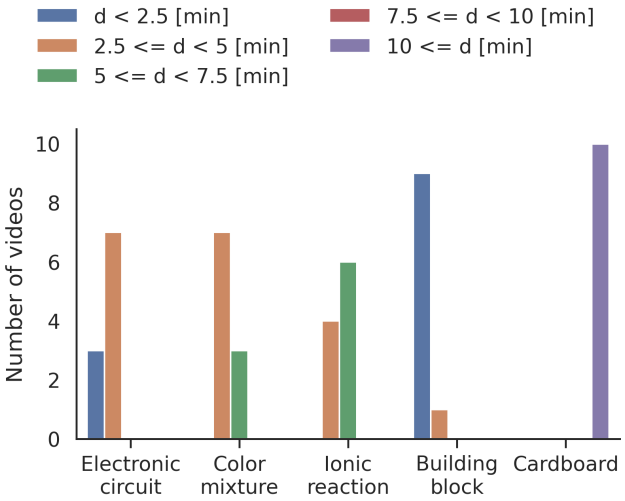


図 5: 映像長（分）の分布．

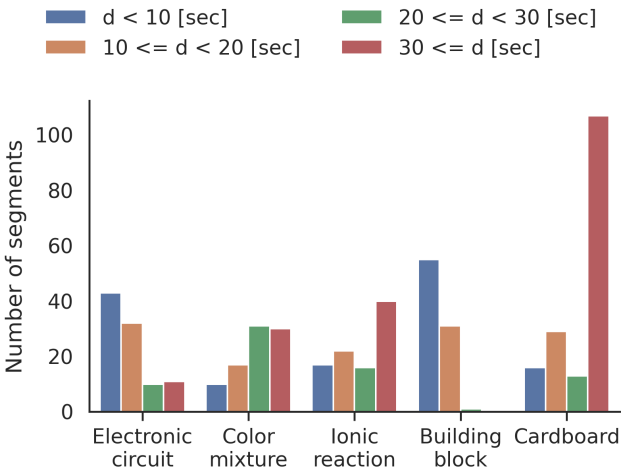


図 6: 区間長（秒）の分布．

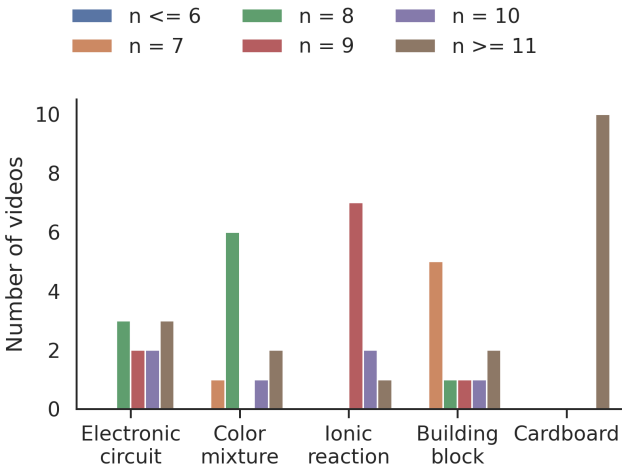


図 7: 区間数の分布．

A　付録

A.1　データセットの構築

マイクロ QR コード　名称を埋め込んだマイクロ QR コードを物体に添付する．作業誤りの中でも物体の誤りを検出するには，操作中の物体を認識する必要がある．しかし，2.1 節で挙げた作業で用いる物体の中には，小さかったり外観が類似していて判別が困難なものがある．こうした物体を物体検出モデルで認識するのは困難であり，マイクロ QR コードを物体に添付する．なお，本稿の実験では，マイクロ QR コードの認識を使用していない．認識結果の作業誤りの検出への活用は，今後の課題である．

映像の統計情報　各映像の映像長，区間数，区間長の分布を，作業ごとに分けて示す．図 5 に示すように，映像長の分布は作業によって異なっている．積み木ではほとんどの映像が 2.5 分以下だが，段ボール工作では全て 10 分以上である．図 6 に示すように，区間長の分布も作業によって異なっている．積み木や電気回路では 10 秒以下の区間が最も多いが，ダンボール工作では 30 秒以上の区間が最も多いかつ大半を占めている．図 7 に示すように，区間数の分布も作業によって異なっている．光の混色，イオン反応，積み木，段ボール工作では，大半の映像で特定の区間数を含む．こうした傾向は，主に手順書の手順数や手順の複雑さの違いに由来すると考えられ，2.4 節で示した手順書の多様性が，映像の多様性にもつながっている．映像には作業間で多様性があり，さらに同じ作業内でも映像同士が異なっていることがわかる．なお，2.4 節にて，映像の統計情報の各作業における平均値を示した．