Egocentric Biochemical Video-and-Language Dataset

Taichi Nishimura[1] Kojiro Sakoda[1]
[1]Graduate School of Informatics, Kyoto University

Atsushi Hashimoto[2] Yoshitaka Ushiku[2]
[2]OMRON SINIC X Corporation

Natsuko Tanaka[3] Fumihito Ono[3]
[3]Osaka Medical and Pharmaceutical University

Hirotaka Kameko[4] Shinsuke Mori[4]
[4]Academic Center for Computing and Media Studies, Kyoto University

Abstract

This paper proposes a novel biochemical video-and-language (BioVL) dataset, which consists of experimental videos, corresponding protocols, and annotations of alignment between events in the video and instructions in the protocol. The key strength of the dataset is its user-oriented design of data collection. We imagine that biochemical researchers easily take videos and share them for another researcher's replication in the future. To minimize the burden of video recording, we adopted an unedited first-person video as a visual source. As a result, we collected 16 videos from four protocols with a total length of 1.6 hours. In our experiments, we conduct two zero-shot video-and-language tasks on the BioVL dataset. Our experimental results show a large room for improvement for practical use even utilizing the state-of-the-art pre-trained video-and-language joint embedding model. We are going to release the BioVL dataset. To our knowledge, this work is the first attempt to release the biochemical video-and-language dataset.

1. Introduction

Science faces a replication crisis. As reported by [1], in wet-lab research (e.g., biochemistry and life science), more than 80% of researchers have failed to reproduce another scientist's experiments. Video-and-language techniques would break through this problem. For example, given recorded videos and protocols, an automatic system can create multimedia protocols, which align instructions in the protocols and events in the videos. They help researchers reproduce experiments by providing quantitative (e.g., time and quantity) and qualitative information (e.g., shapes and colors of samples). Furthermore, video captioning [4] generates protocols only from recorded videos, saving the researcher's effort of writing them.

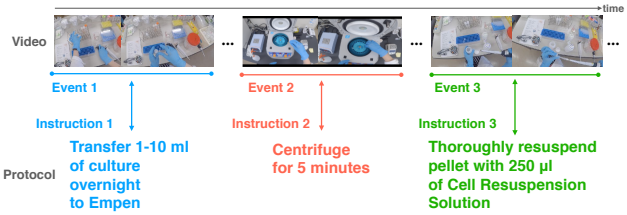

Figure 1. One of the samples of the BioVL dataset. Recorded experimental videos have annotations of an alignment between events in the video and instructions in the protocol.

As the first step towards this goal, this paper proposes a novel Biochemical Video-and-Language (BioVL) dataset (Figure 1), which consists of experimental videos, text protocols, and annotations of alignment between events in the videos and instructions in the protocols. The key strength of the dataset is its user-oriented design of data collection. We imagine that biochemical researchers easily take videos and share them for another researcher's replication in the future. To minimize the burden of video recording, we adopted an unedited first-person video as a visual source, referring to [3], without additional cameras or sensors [8, 12]. As a result, we collected 16 videos from four protocols with a total length of 1.6 hours.

Using the BioVL dataset, we conduct two zero-shot video-and-language tasks for practical applications: instruction-to-event retrieval and instruction-video alignment. Our experimental results show a large room for improvement for practical use even using the state-of-the-art pre-trained video-and-language joint embedding model. The BioVL dataset will be available online only for research purposes; to our knowledge, this work is the first attempt to release the biochemical video-and-language dataset on the web.

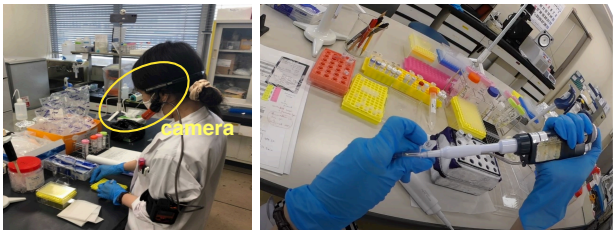

Figure 2. (a) and (b) shows the recording studio of experiments and the view from the equipped first-person camera, respectively.

2. Related Work

2.1. Video-and-Language How-to Dataset

As the biochemical domain is one of the how-to domains (e.g., cooking and assembling furniture), we introduce vision-and-language how-to datasets. The largest dataset is the Howto100M dataset [7], which consists of 100M instructional videos with their narration transcriptions in various domains. Several works [7, 6] reported that pre-training video-and-language models on this dataset contributes to the performance improvement on various downstream tasks. We use one of the models [7] pre-trained on this dataset in our experiments.

In contrast to the significant progress in the how-to domains, little work targets the biochemical domain. The only two studies [8, 9] previously targeted the biochemical domain, tackling an unsupervised alignment problem of instructions in the protocols and events in the video. Unfortunately, their benchmark datasets are not available online, and the other researchers cannot follow their research. To address this issue, we are going to release the BioVL dataset only for research purposes.

2.2. Egocentric How-to Dataset

Recently, much attention has been paid to an egocentric how-to dataset that consists of recorded videos with linguistic annotations (e.g., narration) [3]. The BioVL dataset will be a biochemistry version of the dataset; it will be a benchmark for biochemical video-and-language tasks, such as cross-modal retrieval [5], video-and-language alignment [2], and video captioning [4]. Moreover, it also would be a valuable resource for practical applications, such as robot imitation learning from human demonstration [10] and virtual assistant for researchers to reproduce experiments.

3. BioVL Dataset

This section describes how to collect videos, the annotation processes, and the statistics of the BioVL dataset.

Table 1. An annotation example of PCR.

Instruction	start	end
add sterile distilled water	30	45
add primer1	64	99
add primer2	106	130
add template	149	173
add primeSTAR	190	238
set in DNA engine	260	266

3.1. Video Recording

Participant. We asked one researcher (1 female) for our video collection. During experiments, the researcher put on a headset that fixes a wearable first-person camera[1] (Figure 2). Note that the headset is light enough for researchers to concentrate on their experiments.

Experiment targets. We chose the basic well-known four experiments that have well-established protocols in the biochemical domain: miniprep, PCR, DNA extraction, and making an agarose gel. We took four videos per experiment, thus finally got 16 videos in total[2]. Only DNA extraction has two different methods: Phenol-chloroform extraction and Ethanol precipitation. We took videos two times per method.

Data processing. In a few experiments, the researcher needed to wait while executing specific instructions (e.g., centrifuge samples). During the waiting time, the researcher put off the headset, leaving the camera on. We manually trimmed such waiting times because they are not related to any instructions.

3.2. Annotation

We hired two annotators to align events in the videos and instructions in the protocols; one annotator for the major effort and the other for verification, following [13]. While the former annotator is a non-expert, the latter is an expert annotator. First, the annotators split the protocol sentences into instructions by actions; for example, the sentence "Invert 4 times to mix and add 10 μl of Alkaline Protease Solution." was split into two instructions "Invert 4 times to mix" and "add 10 μl of Alkaline Protease Solution." Then, the annotators watched a video and annotated events by determining start/end times (seconds) for each instruction. In this annotation phase, the latter expert annotator validated events based on the annotation from the former annotator and corrected them if there are any mistakes. We cannot compute the agreement rate of event annotation because we only saved the latter's annotation. In the future, we are going to evaluate the annotation quality by asking other experts to annotate events independently. Table 1 shows an example file of our annotation.

[1] We use Panasonic HX-A500.

[2] The researcher followed the same protocol per experiment. They could read it during the experiments.

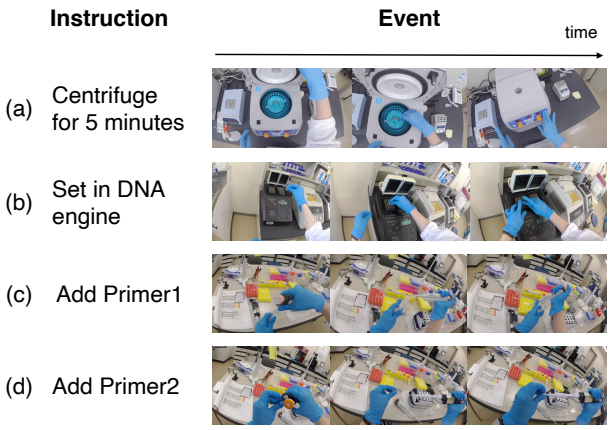

Figure 3. Visual distinctiveness for different instructions.

Table 2. Text-side statistics of the BioVL dataset. This table reports average values and standard deviation. Note that in agarose gel and miniprep, the researcher follows the same protocols but sometimes skips a few instructions depending on the situation. Therefore standard deviation of #instructions becomes over 0.

		#instructions	#words/#instructions
DNA	Phenol chloroform	4.0 (±0.0)	6.0 (±1.9)
	Ethanol	9.0 (±0.0)	4.9 (±2.9)
PCR		6.0 (± 0.0)	3.0 (±1.0)
Agarose gel		10.3 (±0.4)	4.7 (±2.4)
Miniprep		28.2 (±0.4)	6.4 (±2.5)

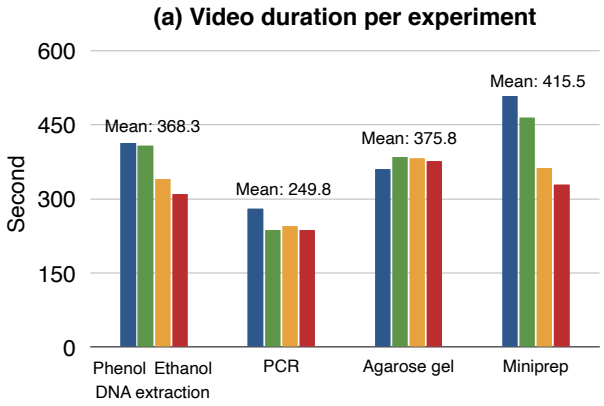

Figure 4. Video duration per experiment.

Visual distinctiveness of different instructions. Figure 3 shows several annotated events for different instructions. This indicates that some instructions are distinctive visually (see (a) and (b)) and others are not (see (c) and (d)). To distinguish them, we have to collect detailed object information by annotating bounding boxes to the video frames.

3.3. Statistics

We finally got 16 videos for four protocols with annotations of alignment between instructions and video events.

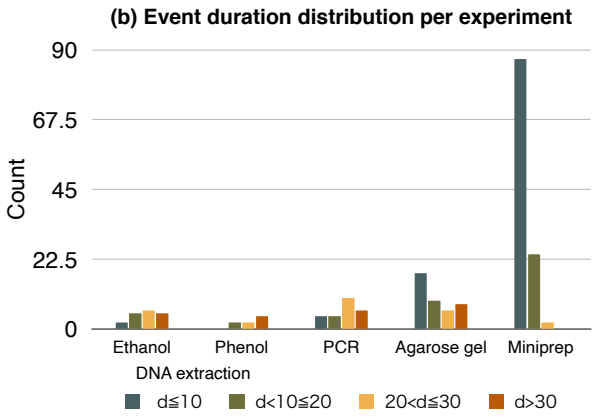

Figure 5. Distribution of the annotated event duration per experiment. d means duration (second).

Table 3. Results of the zero-shot instruction-to-event retrieval.

	MedR	R@1	R@3	R@5
Random	28.8	1.9	5.8	**7.5**
VLE [7] (w/ Howto100M)	**27.0**	**2.0**	**5.9**	5.9

We here discuss the statistics of the BioVL dataset from the text- and video-side. From the text-side, Table 2 indicates a wide range of the number of instructions between experiments; the longest experiment is miniprep while the shortest is Phenol chloroform extraction. From the video-side, while the longest duration is miniprep, the shortest is PCR (see Figure 4). Figure 5 further investigates the distribution of the annotated event duration, showing that a large portion of instructions is short in miniprep; 77% (=87/113) instructions are completed in 10 seconds. These statistics conclude that the BioVL dataset covers both long and short experiments in the biochemical domain.

4. Experiment

We conduct two experiments on the BioVL dataset: (1) instruction-to-event retrieval task and (2) instruction-video alignment task. Due to the limited dataset size, it is infeasible to train a deep model using the BioVL dataset as with the other canonical video-and-language approaches [7, 6]. To this end, we conduct these experiments on the zero-shot settings based on the pre-trained **V**ideo-and-**L**anguage **E**mbedding model, VLE [7]. This model is pre-trained on the Howto100M dataset [7], achieving high performance on various video-and-language tasks[3].

4.1. Zero-Shot Instruction-to-Event Retrieval

One of the promising applications is cross-modal retrieval. Given an instruction as a query, the video-and-

[3]We are going to try the state-of-the-art video-and-language model (e.g., MIL-NCE[6]) in the future.

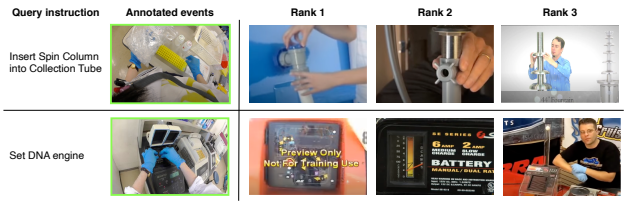

Figure 6. Retrieved results on the howto100M dataset for query instructions from the BioVL dataset.

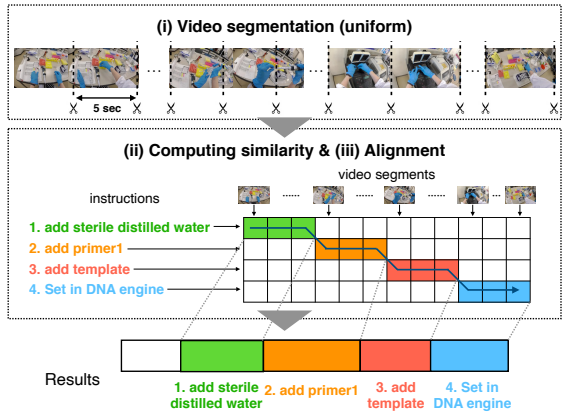

Figure 7. An overview of the alignment method.

language model embeds it and computes the cosine similarly as a score between the query vector and all the 204 event vectors. Then, we sort scores with events in descending order and calculate the median rank (MedR) and recall rate at the top K (R@K). The median rank indicates the median ranking of retrieved corresponding events, hence lower is better; in contrast, R@K represents the percentage of all the instruction queries where the corresponding event is retrieved in the top K, hence higher is better. Note that because we took four videos per protocol, instruction-event pairs are one-to-many. Therefore we compute these metrics by regarding events that have the same protocol's instructions to query as correct events.

Results. Table 3 shows the results of the instruction-to-event retrieval. Even though we use the state-of-the-art video-and-language model, the results are competitive to the random baseline. One of the reasons for this poor performance is that the howto100M dataset does not cover the biochemical domain. To clarify this, we tried to input several instructions in the BioVL dataset to retrieve events from the Howto100M dataset (Figure 6), showing that the retrieved events are not related to the input instructions. One of the solutions to this problem is to collect biochemical videos on the web to train the video-and-language model.

4.2. Zero-Shot Instruction-Video Alignment

Another application is cross-modal alignment. To the best of our knowledge, no work tackled this problem in the zero-shot setting. Therefore, we proposed an alignment method (Figure 7), which consists of three processes: (i) a video is uniformly segmented in $t = 5$ second interval, (ii) the video-and-language model computes cosine similarity between video segments and instructions as scores, (iii) the best alignment are obtained based on the Needleman–Wunsch algorithm [11] by filling each cell in the following DP table:

$$table[i][j] = \begin{cases} 0 & (i = 0 \vee j = 0) \\ \max \left\{ \begin{array}{l} table[i-1][j-1] + score, \\ table[i][j-1] \end{array} \right\} & (\text{otherwise}) \end{cases}$$

For evaluation, we compute mean IoU between selected segments and ground-truth events, following traditional video-and-language alignment evaluation [2].

Table 4. Results of the zero-shot instruction-video alignment.

	mIoU
Uniform	28.6
VLE [7] (w/ Howto100M)	**30.4**

Results. Table 4 shows the results of the instruction-video alignment evaluation. As with the instruction-to-event retrieval, the model is competitive to the uniform baseline, which segments a video in a uniform interval $\frac{T}{L}$, where T and L represents video duration and the number of instructions, respectively. To improve the performance, we are going to try weakly-supervised video-and-language methods [2] by increasing videos/protocols in the future.

5. Conclusion

This paper proposed the novel BioVL dataset, consisting of recorded experimental videos, text protocols, and annotations of alignment between video events and protocol instructions. As a first trial, we collected 16 videos from four protocols with a total length of 1.6 hours. We conducted two zero-shot video-and-language experiments on the BioVL dataset: (1) instruction-to-event retrieval and (2) instruction-video alignment. Our experimental results show a large room for improvement for practical use even using the state-of-the-art pre-trained video-and-language joint embedding model. Our dataset is the first attempt to release a video-and-language dataset for the biochemical domain. We hope that the BioVL dataset encourages computer vision and natural language processing researchers to try biochemical video-and-language problems in the future.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP21J20250 and JP20H04210, and partially supported by JP21H04910, JP17H06100, JST-Mirai Program Grant Number JPMJMI21G2, and JST ACT-I Grant Number JPMJPR17U5.

References

[1] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533:452–454, 2016.

[2] Piotr Bojanowski, Remi Lajugie, Edouard Grave, Francis Bach, Ivan Laptev, Jean Ponce, and Cordelia Schmid. Weakly-supervised alignment of video with text. In *Proc. ICCV*, pages 4462–4470, 2015.

[3] 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 *Proc. ECCV*, pages 753–771, 2018.

[4] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proc. ICCV*, pages 706–715, 2017.

[5] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: efficient text-to-visual retrieval with transformers. In *Proc. CVPR*, pages 9826–9836, 2021.

[6] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proc. CVPR*, pages 9879–9889, 2020.

[7] 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 *Proc. ICCV*, pages 2630–2640, 2019.

[8] Iftekhar Naim, Young Song, Qiguang Liu, Henry Kautz, Jiebo Luo, and Daniel Gildea. Unsupervised alignment of natural language instructions with video segments. In *Proc. AAAI*, pages 1558–1564, 2014.

[9] Iftekhar Naim, Young C. Song, Qiguang Liu, Liang Huang, Henry Kautz, Jiebo Luo, and Daniel Gildea. Discriminative unsupervised alignment of natural language instructions with corresponding video segments. In *Proc. NAACL*, pages 164–174, 2015.

[10] Ashvin Nair, Dian Chen, Pulkit Agrawal, Phillip Isola, Pieter Abbeel, Jitendra Malik, and Sergey Levine. Combining self-supervised learning and imitation for vision-based rope manipulation. In *Proc. ICRA*, pages 2146–2153, 2017.

[11] Saul B. Needleman and Christian D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Jornal of Molecular Biology*, 48:443–453, 1970.

[12] Marcus Rohrbach, Anna Rohrbach, Michaela Regneri, Sikandar Amin, Mykhaylo Andriluka, Manfred Pinkal, and Bernt Schiele. Recognizing fine-grained and composite activities using hand-centric features and script data. *IJCV*, 119:1–28, 2015.

[13] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *Proc. AAAI*, pages 7590–7598, 2018.