Building a Video-and-Language Dataset with Human Actions for Multimodal Logical Inference

Riko Suzuki[1]
suzuki.riko@is.ocha.ac.jp

Hitomi Yanaka[2]
hyanaka@is.s.u-tokyo.ac.jp

Koji Mineshima[3]
minesima@abelard.flet.keio.ac.jp

Daisuke Bekki[1]
bekki@is.ocha.ac.jp

[1]Ochanomizu University, Tokyo, Japan
[2]The University of Tokyo, Tokyo, Japan
[3]Keio University, Tokyo, Japan

Abstract

This paper introduces a new video-and-language dataset with human actions for multimodal logical inference, which focuses on intentional and aspectual expressions that describe dynamic human actions. The dataset consists of 200 videos, 5,554 action labels, and 1,942 action triplets of the form ⟨subject, predicate, object⟩ that can be translated into logical semantic representations. The dataset is expected to be useful for evaluating multimodal inference systems between videos and semantically complicated sentences including negation and quantification.

1 Introduction

Multimodal understanding tasks (Johnson et al., 2017; Suhr et al., 2017, 2019) have attracted rapidly growing attention from both computer vision and natural language processing communities, and various multimodal tasks combining visual and linguistic reasoning, such as visual question answering (Antol et al., 2015; Acharya et al., 2019) and image caption generation (Vinyals et al., 2015), have been introduced. With the development of the multimodal structured datasets such as Visual Genome (Krishna et al., 2017), recent studies have been tackling a complex multimodal inference task such as Visual Reasoning (Suhr et al., 2019) and Visual-Textual Entailment (VTE) (Suzuki et al., 2019; Do et al., 2020), a task to judge if a sentence is true or false under the situation described in an image.

The recently proposed multimodal logical inference system (Suzuki et al., 2019) uses first-order logic (FOL) formulas as unified semantic representations for text and image information. The FOL formulas are structured representations that capture not only objects and their semantic relationships in images but also those complex expressions including negation, quantification, and numerals. When we consider extending the logical inference system between texts and images to that between texts and videos, it is necessary to handle the property of video information: there are dynamic expressions to capture human actions and movements of things in videos more than in images.

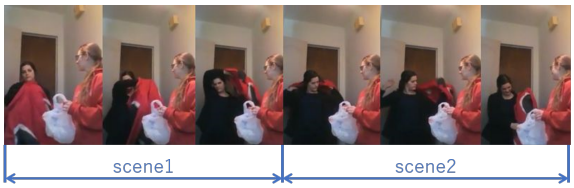

Figure 1: Inference example between a video and sentences. The description of this video is: *The woman tried to put on her outerwear though she could not, because its zipper was not open completely.*

As an example, consider a video-and-language inference example in Figure 1. This video consists of SCENE1, where the sentence *The woman puts on her outerwear* is true, and SCENE2, where the sentence *The woman takes off her outerwear* is true. Note that the entire video represents richer information as expressed by the sentence *the woman tries to put on her outerwear*. To judge whether this sentence is true, it is not enough to simply combine two actions, *putting on outerwear* and *taking off outerwear*. To capture this dynamic aspect of human action, it is necessary to take into account the information expressed by intentional phrases such as *trying to put on outerwear*.

Towards such a complex multimodal inference between video and text, we build a new Japanese video-and-language dataset with human actions. We annotate videos with action labels written in triplets of the form ⟨subject, predicate, object⟩, where object can be empty (indicated by ϕ). Action labels contain not only basic expressions such as ⟨person, run, ϕ⟩ and ⟨person, hold, cup⟩,

Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR), pages 102–107

but also expressions including intentional phrases such as ⟨person, try to eat, food⟩. An advantage of using triplets ⟨subject, predicate, object⟩ is that a triplet itself can serve as the semantic representation of a video and can be translated into logical formulas (see Section 3). This paper introduces a method to create a video-and-language dataset involving aspectual and intentional phrases. We collect a preliminary dataset labeled in Japanese for human actions. We also analyze to what extent our dataset contains various aspectual and intentional phrases. Our dataset will be publicly available at https://github.com/rikos3/HumanActions.

2 Related Work

There have been several efforts to create human action video datasets in the field of computer vision. Charades (Sigurdsson et al., 2016) contains 9,848 videos of daily activities annotated with free-text descriptions and action labels in English. Charades STA (Gao et al., 2017) is a dataset built by adding sentence descriptions with start and end times to the Charades dataset. For Japanese video datasets, STAIR Actions (Yoshikawa et al., 2018) is a dataset that consists of 63,000 videos with action labels. Each video is about 5 seconds and has a single action label from 100 action categories. Action Genome (Ji et al., 2020) is a large-scale video dataset built upon the Charades dataset, which provides action labels and spatio-temporal scene graphs.

VIOLIN (Liu et al., 2020) introduces a multimodal inference task between text and videos: given a video with aligned subtitles as a premise, paired with a natural language hypothesis based on the video content, a model needs to judge whether or not the hypothesis is entailed by the given video. The VIOLIN dataset mainly focuses on conversation reasoning and commonsense reasoning, and the dataset contains videos collected from movies or TV shows.

Compared to the existing datasets, our dataset is distinctive in that action labels are written in structured representations ⟨subject, predicate, object⟩ and contain various expressions such as *continue to eat* and *try to close* that support complex inference between videos and texts.

3 Semantic Representations of Videos

Suzuki et al. (2019) proposed FOL formulas as semantic representations of text and images. They use the formulas translated from FOL structures for images to solve a complex VTE task. We extend this idea to semantic representations of videos.

FOL structures (also called first-order *models*) are used to represent semantic information in images (Hürlimann and Bos, 2016). An FOL structure for an image is a pair (D, I) where D is a domain consisting of all the entities occurring in the image, and I is an interpretation function that describes the attributes and relations holding of the entities in the image (Suzuki et al., 2019).

To extend FOL structures for images to those for videos, we add to FOL structures a set of scenes $S = \{s_1, s_2, \ldots, s_n\}$ that makes up a video, ordered by the temporal precedence relation. This structure may be considered as a possible world model for standard temporal logic (Venema, 2017; Blackburn et al., 2002). Thus, a video is represented by (S, D, I), where S is a set of scenes linearly ordered by the temporal precedence relation, D is a domain of the entities, which is constant in all scenes, and I is an interpretation function that assigns attributes and relations to the entities in each scene. We assign personal IDs $(\mathsf{d}_1, \mathsf{d}_2, \ldots, \mathsf{d}_n)$ to people appearing in each scene. Since the purpose of our dataset is to label human actions, we assign IDs to people, but not to non-human objects.

To facilitate the annotation of the attributes and relations holding of the entities in each scene, we use triplets of the form ⟨subject, predicate, object⟩ given to each scene $\mathsf{s_i}$ as action labels, where object may be empty. This form itself can be seen as a semantic representation of videos. Furthermore, it can also be translated into an FOL formula, in a similar way to the standard translation of modal logic to FOL (Blackburn et al., 2002). The following examples show a translation from triplets in scenes into FOL formulas.

(1) $\mathsf{s}_1 : \langle \mathsf{d}_1, \mathsf{run}, \phi \rangle$
$\Rightarrow \mathsf{run}(\mathsf{s}_1, \mathsf{d}_1)$

(2) s_2: $\langle \mathsf{d}_1, \mathsf{hold}, \mathsf{pillow} \rangle$
$\Rightarrow \exists x(\mathsf{pillow}(\mathsf{s}_2, x) \wedge \mathsf{hold}(\mathsf{s}_2, \mathsf{d}_1, x))$

Here each predicate has an additional argument for a scene variable. (1) means that the entity d_1 runs in scene s_1; (2) means that the entity d_1 holds a pillow in scene s_2.

Each triplet can be translated into an FOL formula by using this method and thus serve as a se-

Figure 2: Example video for the action of *touching someone's shoulder* from the Charades dataset.

mantic representation of a video usable in the semantic parser and inference system for the VTE task presented in Suzuki et al. (2019). Though it is left for future work, the dataset in which each scene of a video is annotated with triplets will be useful to evaluate the VTE system for videos.

4 Data Collection

4.1 Video Selection

We selected videos from the test set of the Charades dataset (Sigurdsson et al., 2016). The Charades dataset contains videos drawing daily activities in a room such as *drinking from a cup*, *putting on shoes*, and *watching a laptop or something on a laptop*. Each video is collected via crowdsourcing: workers are asked to generate the script that describes daily activities and then to record a video of that script being acted out.

We select videos where multiple persons appear from the Charades test set to cover various actions within human interaction such as *touching someone's shoulder* or *handing something*. These actions are expected to be described in expressions involving various linguistic phenomena. To collect videos where multiple persons appear, we selected 200 videos whose descriptions include phrases *another person*, *another people*, and *they*. Figure 2 shows a video example involving human interaction.

4.2 Annotation

We annotate each video with ⟨subject, predicate, object⟩ triplet format as action labels that represent human-object activities. We also annotate each action label with a start and end time to locate the activity accurately. We ask two workers to freely write predicates and object names that describe human activities to collect various expressions. Using this format the workers can freely decide the span of each scene and thus annotate a video with action labels more easily and flexibly. In Section 4.5 below, we will explain how to convert the triplet action format with start and end times to FOL structures extended with scenes as presented in Section 3.

Subject We assign personal IDs ($\mathsf{d}_1, \mathsf{d}_2, \mathsf{d}_3, \ldots$) to people in order of appearance in the video. If multiple persons appear for the first time in the same scene, we assign personal IDs to people appearing in order from left to right.

Predicate In a triplet, predicate contains various expressions such as aspectual and intentional phrases for describing dynamic human actions in videos, those phrases that do not usually appear in captions for static images. The following examples show characteristic predicates of videos.

- predicates for utterance and communication (e.g. *speak*, *talk*, *tell*, *ask*, *listen*)
- predicates for intention and attitude (e.g. *try to eat*, *try to close*).
- aspectual predicates (e.g. *start talking*, *continue to eat*)

We allow workers to use not only a transitive or intransitive verb but also verb phrases for predicates such as *try to V* and *continue to V* to collect diverse aspectual and intentional phrases.

Object The object in a triplet contains an object name or personal ID. If the item in predicate is an intransitive verb, object is empty. For instance, in Figure 3, the object for the predicate hold is pillow and the object is empty for the predicate run.

Figure 3: *A man is running while holding a pillow.* Action labels are ⟨d_1, hold, pillow⟩ and ⟨d_1, run, ϕ⟩

4.3 Validation

In this work, we ask three workers to either annotate or merge action labels. All of the workers are native speakers of Japanese. We merge and confirm action labels in the following steps: (1) merge action labels made by two workers and arrange them in ascending order of start times, (2) watch videos by three workers to see if an action label is correct, and (3) if action labels duplicate, select one action label.

Regarding duplicated action labels, the labels and their start and end time are determined according to the agreement of three workers. Consider the following duplicate case.

Dataset	Videos	Average time (sec)	Average of action labels	Action categories	English	Japanese
Charades (Sigurdsson et al., 2016)	9848	30	6.8	157	✓	
ActionGenome (Ji et al., 2020)	9848	30	170	157	✓	
STAIR Actions (Yoshikawa et al., 2018)	102462	5-6	1.0	100	✓	✓
Ours	200	30	27.77	**1942**		✓

Table 1: A comparison of our dataset with existing datasets

Predicate	Freq.	Examples
Utterance	138 (2.49%)	話す/*talk*(102), 喋る/*speak*(20), 話しかける/*address*(11), 声を出す/*speak*(3), 歌う/*sing*(1), 話しかけられる/*be spoken*(1)
Intention/ Attitude	51 (0.98%)	閉めようとする/*try to close*(7), 飲もうとする/*try to drink*(6), 持とうとする/*try to hold*(3), 置こうとする/*try to put*(3), 切ろうとする/*try to cut*(2), 動かそうとする/*try to move*(2), 食べるふりをする/*pretend to eat*(2), 外そうとする/*try to remove*(2), 着ようとする/*try to put on*(2)
Aspect	8 (0.15%)	止める/*stop*(4), 食べ続ける/*continue to eat*(1), かけるのを止める/*stop to hang*(1), 組み立て続ける/*continue to build*(1), 覗き続ける/*continue to peep*(1)

Table 2: Predicates for utterance, intention and aspect

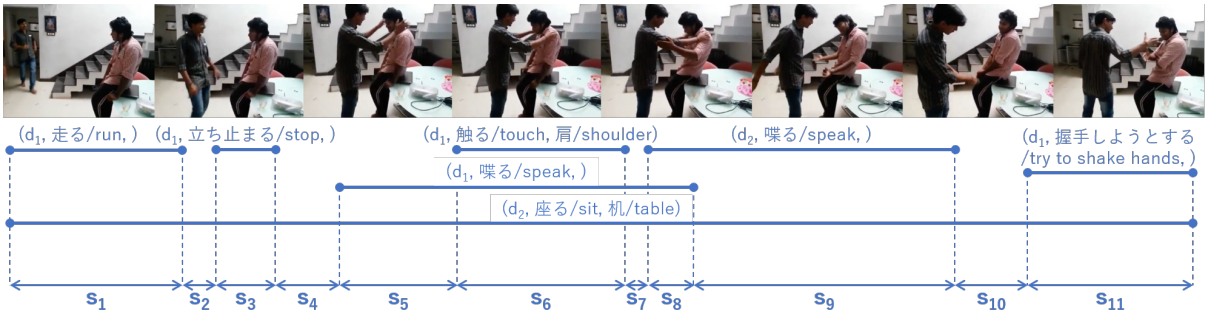

Figure 4: Annotation example of a video labeled with various types of predicates. Here $s_1, \ldots, s_{11}$ are scenes linearly ordered by the temporal precedence relation.

(σ_1) 0:10-0:13 $\langle \mathsf{d}_1, \mathsf{hold}, \mathsf{clothes} \rangle$
(σ_2) 0:11-0:14 $\langle \mathsf{d}_1, \mathsf{hold}, \mathsf{clothes} \rangle$
(σ_3) 0:11-0:15 $\langle \mathsf{d}_1, \mathsf{hold}, \mathsf{outerwear} \rangle$

In this case, (σ_1) and (σ_2) are duplicates in that subject, predicate, and object are the same while the start time and end time are different. If the third worker judges that (σ_2) is more adequate than (σ_1), we merge (σ_1) and (σ_2) and obtain the action labels below.

(σ_1') 0:10-0:14 $\langle \mathsf{d}_1, \mathsf{hold}, \mathsf{clothes} \rangle$
(σ_2') 0:11-0:15 $\langle \mathsf{d}_1, \mathsf{hold}, \mathsf{outerwear} \rangle$

4.4 Collection Statistics

Table 1 shows that despite its size, our dataset contains more action categories than other previous datasets. About 65% of total action labels are action labels that appear only once. This indicates that there are a wide variety of expressions.

The dataset contains characteristic expressions of videos such as *walk*, *talk*, and *stop walking*. Table 2 shows the frequency and examples of three types of predicates, i.e., utterance, intentional, and aspectual predicates. The distribution of characteristic predicates of videos in our dataset was: 2.49% predicates for utterance, 0.98% predicates for intention and attitude, and 0.15% aspectual predicates. One possible reason for the low frequency of aspectual predicates is that Charades contains 30-second videos, which might be too short to describe multiple actions involving aspectual phrases. It would be expected to increase the number of aspectual predicates if we annotate

Action label	Freq.	Rate(%)
歩く/*walk*	288	5.19
立つ_床/*stand_floor*	221	3.98
立ち止まる/*stop walking*	102	1.84
立つ/*stand*	96	1.73
見る/*see*	81	1.46
話す/*talk*	81	1.46
笑う/*laugh*	71	1.28
食べる_食べ物/*eat_food*	54	0.97
飲む_飲み物/*drink_beverage*	48	0.86
持つ_コップ/*hold_cup*	47	0.85

Table 3: Top 10 frequent action labels. Action labels are written in form of predicate_object or predicate.

longer videos such as the VIOLIN dataset (Liu et al., 2020), which is left for future work. The number of overlaps of action categories between ours and STAIR Actions (Yoshikawa et al., 2018) is 28. These results indicate that our dataset contains more diverse action categories compared to other datasets.

Table 3 shows frequent action labels in our dataset. Our dataset contains not only predicates for utterance, intention, and aspect, but also punctual verbs (e.g. *stop walking* and *turn on*) and durative verbs (e.g. *sit* and *wait*).

4.5 Conversion to FOL structures

The triplet action forms with start and end points used in the annotation can be converted to FOL structures extended with scenes presented in Section 3. In the extended FOL structures, each scene is linearly ordered by the temporal precedence relation and is uniquely characterized by the set of all the attributes and relations holding in it.

As an illustration, consider the example in Figure 4. In this case, we can separate the entire video into 11 scenes as shown in Figure 4. Accordingly, in the extended FOL structure, we have $S = \{s_1, \ldots, s_{11}\}$. Here the first scene, s_1, consists of the following: the predicate run holds of the entity d_1, the predicate sit holds of the pair (d_2, x_1) where x_1 is an entity which is a table. In terms of the interpretation function I relativized to a scene, we have $I_{s_1}(\mathsf{run}) = \{\mathsf{d}_1\}, I_{s_1}(\mathsf{sit}) = \{(\mathsf{d}_2, x_1)\}$ and $I_{s_1}(\mathsf{table}) = \{x_1\}$. Similarly, we can extend the interpretation function I to the other scenes.

While the triplet format is suitable for the annotation of various action labels, the semantic representation in the form of FOL structures with scenes can be directly used in model checking and theorem proving for the VTE system developed in Suzuki et al. (2019). Our annotation format is flexible enough to be adapted in such applications.

5 Conclusion

We introduce a video-and-language dataset with human actions for multimodal inference. We annotate human actions in videos in the free format and collect 1,942 action categories for 200 videos. Our dataset contains various action labels for videos, including those predicates characteristic of videos such as predicates for utterance, predicates for intention and attitude, and aspectual predicates. In future work, we analyze recent action recognition models using Action Genome (Ji et al., 2020) with our dataset. We will also work on building a multimodal logical inference system between texts and videos.

Acknowledgment

This work was partially supported by JST CREST Grant Number JPMJCR20D2, Japan. Thanks to the anonymous reviewers for helpful comments. We would also like to thank Mai Yokozeki and Natsuki Murakami for their contributions.

References

Manoj Acharya, Kushal Kafle, and Christopher Kanan. 2019. TallyQA: Answering complex counting questions. In *The Association for the Advancement of Artificial Intelligence (AAAI2019)*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision*.

Patrick Blackburn, Maarten de Rijke, and Yde Venema. 2002. *Modal Logic*. Cambridge University Press.

Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. 2020. e-SNLI-VE-2.0: Corrected visual-textual entailment with natural language explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. TALL: temporal activity localization via language query. In *IEEE International Conference on Computer Vision*, pages 5277–5285, Venice, Italy. IEEE Computer Society.

Manuela Hürlimann and Johan Bos. 2016. Combining Lexical and Spatial Knowledge to Predict Spatial Relations between Objects in Images. In *Proc. of the Workshop on Vision and Language*.

Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. 2020. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. 2020. Violin: A large-scale dataset for video-and-language inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10900–10910.

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 510–526, Amsterdam, Netherlands. Springer.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada. Association for Computational Linguistics.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

Riko Suzuki, Hitomi Yanaka, Masashi Yoshikawa, Koji Mineshima, and Daisuke Bekki. 2019. Multimodal logical inference system for visual-textual entailment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 386–392, Florence, Italy. Association for Computational Linguistics.

Yde Venema. 2017. *Temporal Logic*, chapter 10. John Wiley and Sons, Ltd.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A neural image caption generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164.

Yuya Yoshikawa, Jiaqing Lin, and Akikazu Takeuchi. 2018. STAIR actions: A video dataset of everyday home actions. *CoRR*, abs/1804.04326.