Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4664–4670
Marseille, 11–16 May 2020

Video Caption Dataset for Describing Human Actions in Japanese

Yutaro Shigeto, Yuya Yoshikawa, Jiaqing Lin, Akikazu Takeuchi
STAIR Lab, Chiba Institute of Technology
Tsudanuma 2-17-1, Narashino, Chiba, Japan
{shigeto, yoshikawa, lin, takeuchi}@stair.center

Abstract

In recent years, automatic video caption generation has attracted considerable attention. This paper focuses on the generation of Japanese captions for describing human actions. While most currently available video caption datasets have been constructed for English, there is no equivalent Japanese dataset. To address this, we constructed a large-scale Japanese video caption dataset consisting of 79,822 videos and 399,233 captions. Each caption in our dataset describes a video in the form of "who does what and where." To describe human actions, it is important to identify the details of a person, place, and action. Indeed, when we describe human actions, we usually mention the scene, person, and action. In our experiments, we evaluated two caption generation methods to obtain benchmark results. Further, we investigated whether those generation methods could specify "who does what and where."

Keywords: video captioning, caption generation, Japanese caption dataset, human action understanding

1. Introduction

Automatic video caption generation is a task that outputs the description, or caption, of an input video (Venugopalan et al., 2015b; Venugopalan et al., 2015a; Yao et al., 2015). Video caption generation has many practical applications such as video searching using natural language queries, natural language video summarization, and use as a communication robot. It can also be useful for visually impaired people.

This paper tackles one of such application, namely automatic human action description generation. To describe human actions, it is necessary to recognize and understand "who does what and where." When we explain human actions, it is important to include details of the person, place, and action. While various researchers have already introduced video caption datasets for describing human actions (Sigurdsson et al., 2016; Krishna et al., 2017; Awad et al., 2018), none of these datasets evaluate those aspects individually.

Another problem is that there are difficulties in captioning resource-poor languages. Most previous research has focused on English caption generation due to the scarcity of resources targeting other languages. Each language is unique in terms of properties such as grammar and multiword expressions, making it difficult to determine how to generate captions in other languages. Conversely, the practical applications of caption generation are common to all languages. Hence there is massive demand for caption generation in languages other than English.

These issues motivated us to develop a video caption dataset for describing human actions in Japanese (Sect. 2). This dataset is based on 79,822 videos collected from STAIR Actions, a dataset for human action recognition (Yoshikawa et al., 2018). Each video has five descriptions on average, resulting in 399,233 captions in total. Each caption specifies "who does what and where," and is written in Japanese. This is the first instance of a Japanese video caption dataset, and is the most extensive dataset available, in relation to existing English caption datasets, although English and Japanese are clearly entirely different languages, and therefore, their statistics are not directly comparable.

			The number of characters			
	Uniq.	Voc.	Mean	Median	Max	Min
PLACE	49,460	5,214	6.3	6.0	60	1
PERSON	73,966	4,383	10.0	10.0	55	1
ACTION	110,926	10,098	11.9	11.0	73	1
sentence	306,116	13,836	30.2	28.0	135	8

Table 1: Our dataset statistics. "Uniq." indicates the number of unique phrases/sentences and "Voc." is the vocabulary size. "sentence" (bottom row) represents statistics for sentences obtained using the template.

In our experiments, we obtained benchmark results for this dataset (Sect. 5), investigating whether captioning methods could specify "who does what and where," in addition to standard generation evaluation such as BLEU, ROUGE, and CIDEr.

Our caption dataset is publicly available on our homepage.[1]

2. Japanese Caption Annotations for STAIR Actions

To construct the video caption dataset, we first collected videos from an existing video dataset and then asked workers to annotate multiple (approximately five) captions for each video, resulting in a dataset of 79,822 videos and 399,233 captions.[2]

2.1. Video Collection

Videos were sourced from STAIR Actions dataset (Yoshikawa et al., 2018); a video dataset for human action recognition. Each video in this dataset contains a single human action from a set of 100 everyday actions (e.g., shaking hands, dancing, and reading a book). The average video length is approximately 5 seconds long,

[1]https://actions.stair.center
[2] We used an annotation service provided by BAOBAB Inc.

(a) Input video (1 fps).

PLACE	PERSON	ACTION
街中 (the city)	青い洋服の男の子 (a boy with blue clothes)	写真を撮っている (is being taken a photo)
屋外 (outdoors)	青い服を着た男性 (a man worn blue clothes)	写真を撮っている (is being taken a photo)
黒い柱のある道路 (the road with black pillars)	水色の服を着た少年 (a boy worn light blue clothes)	怪物のコスプレをした人と写真を撮ってもらっている (is being taken a photo with a person who made a cosplay of a monster)
車と黒い柱のある屋外 (outdoor space with car and black pillars)	金色の仮装をした男性 (a man worn golden costumes)	立って子供を抱えている (is standing and holding a child)
石造りの建物のある歩道 (the pavement with a stone building)	羽のついている金の衣装を着た人 (a person worn gold costumes with wings)	子供と一緒に写真を撮っている (is being taken a photo with children)

(b) Phrase annotations.

Figure 1: An example of (a) an input video and (b) its phrase annotations. A sentence can be obtained by filling in the slots in the format: PLACE で PERSON が ACTION.

with a frame rate of 30 fps. STAIR Actions dataset was thus a good fit with our objective of describing single actions (i.e., "who does what and where").

2.2. Caption Annotation

Human actions can essentially be described in terms of "who does what and where," with action descriptions typically mentioning the scene, person, and the specific action. On this basis, three elements were set as a requirement of our captions.

To annotate the three elements, a question answering annotation procedure was performed. First, we asked workers the following questions about a video:

- Who is present? (PERSON)
- Where are they? (PLACE)
- What are they doing? (ACTION)

In this procedure, acceptable answers were a noun phrase for PERSON and PLACE and a verb phrase for ACTION. Further, we set the following annotation guidelines:

(1) A phrase must describe only what is happening in a video and the things displayed therein.
(2) A phrase must not include one's emotions or opinions about the video.
(3) If one does not know the location, write 部屋 (*room*), 屋内 (*indoor*), or 屋外 (*outdoor*).
(4) If one does not know who the person is, write 人 (*person*).

The phrases obtained were reviewed, and corrected if inaccurate. The annotation work was completed by 125 workers in four months. Figure 1 shows an example of our captions. After phrase annotations were completed, sentences were obtained by complementing Japanese particles で and が:

PLACE で PERSON が ACTION.

Dataset	#videos	#captions
MSVD (Chen and Dolan, 2011)	2k	86k
TACoS ML (Rohrbach et al., 2014)	14k	53k
MSR-VTT (Xu et al., 2016)	10k	200k
Charades (Sigurdsson et al., 2016)	10k	16k
LSMDC (Rohrbach et al., 2017)	118k	118k
ActivityNet (Krishna et al., 2017)	100k	100k
YouCook II (Zhou et al., 2018)	15k	15k
VideoStory (Gella et al., 2018)	123k	123k
TRECVID (Awad et al., 2018)	2k	10k
Ours	80k	399k

Table 2: Video caption datasets.

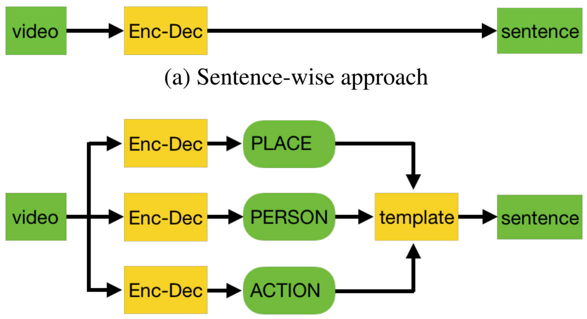
Obviously, this template-based sentence construction does not produce grammatically differing sentences. Since the objective of this research is to specify human actions, the captions may not require complex sentence patterns such as anastrophe and taigendome (a rhetorical device in Japanese); i.e., ending a sentence with a noun. Moreover, the sentences produced were not unnatural.

As a result, we obtained a total of 399,233 sentences. Table 1 shows the statistics for the annotated phrases and the sentences obtained using the template. As the table shows, the unique sentences account for 76.7% of the total. For determining vocabulary size, we used KyTea[3] (Neubig et al., 2011), a morphological analyzer, to tokenize each phrase/sentence into words. In PLACE, the frequency of the terms (部屋, 屋内, and 屋外) was 118,092; i.e., these terms comprise one third of the phrases. There were 27,835 instances of 人; that is, 7% of PERSON phrases.

3. Related Work

Many video caption datasets have been constructed recently, including MSVD (Chen and Dolan, 2011),

[3]http://www.phontron.com/kytea/



(a) Sentence-wise approach

(b) Phrase-wise approach

Figure 2: Overview of two sentence generation approaches. (a) A sentence-wise generation approach generates a sentence directory using a single encoder-decoder model. (b) A phrase-wise approach first generates three phrases (PLACE, PERSON, and ACTION) separately, and then outputs a complete sentence using the template.

Charades (Sigurdsson et al., 2016), ActivityNet (Krishna et al., 2017), and TRECVID (Awad et al., 2018). Table 2 summarizes the video caption datasets most commonly used in video captioning experiments. Apart from MSVD, these datasets only provide English descriptions, and while MSVD contains 15 languages captions besides English, it has a limited number of captions (i.e., 6,245 captions at most) in other languages and none in Japanese language. Differing from MSVD, the dataset described in this paper provides descriptions in Japanese not English.

There are some existing video caption datasets for describing human actions. ActivityNet is a video caption dataset whose main objective is to detect and describe numerous events (human actions) in a long video (180 seconds on average), requiring the ability to recognize the dependencies between human actions. Conversely, each video in our dataset only contains just one action, which is appropriate for our research objective. Charades also provides descriptions of human actions. However, participant details are insufficient, with “a person” appearing frequently in the captions. Each sentence in TRECVID includes four elements of the video: Who, what, where, and when. Our dataset is similar in spirit to the TRECVID dataset but ours is larger. TRECVID contains about 2k videos with 10k captions, while our dataset has approximately 80k videos with 399k captions.

4. Sentence Generation

We evaluated two sentence generation approaches: sentence-wise and phrase-wise approaches. Figure 2 presents an overview of both.

The sentence-wise approach generates a whole sentence using a single encoder-decoder model; a standard approach in the video captioning literature.

The phrase-wise approach uses three encoder-decoders. It first generates PLACE, PERSON, and ACTION, respectively, and then fills in slots in the template (i.e., PLACE で PERSON が ACTION). This approach is a reasonable way of achieving the current objective of generating a description that specifies “who does what and where.” Another advantage to this approach is that the training decoders for phrase generation is easier than for sentence generation. Since phrases are shorter than sentences, it is sufficient for the decoders to target relatively short sequences of words.

In our experiments, we used a multi-modality fusion caption generation method (Jin et al., 2016)—the winning solution of the MSR Video to Language Challenge 2016—as the encoder-decoder model in both sentence-wise and phrase-wise approaches; a method frequently used as a baseline in video captioning experiments. In this method, the encoder (a multilayer feedforward network) transforms multi-modality features into a single vector, and the decoder (a recurrent neural network) generates a sequence of words from the vector.

The output word sequence is chosen by beam search. To eliminate length bias, we used the length normalization presented in Wu et al. (2016).

5. Experiment

We used sentence and phrase generation tasks to evaluate our dataset. The objective of this experiment was to investigate two points: (i) whether the methods can specify “who does what and where” and (ii) the differences between sentence-wise and phrase-wise approaches.

5.1. Experimental Setups

Dataset We randomly split the videos into training (80%), development (10%), and test (10%) sets.

We ran SentencePiece[4] (Kudo and Richardson, 2018), an unsupervised text tokenizer, to segment captions into subwords. We trained the SentencePiece model on a subset of the entire set of captions used to train the captioning methods. The vocabulary size of this model was set to 8,000.

Evaluation Criteria In accordance with the literature (Long et al., 2018; Pan et al., 2017; Gan et al., 2017; Wang et al., 2018; Phan et al., 2017), generated captions were evaluated based on three criteria; BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015). In the evaluation phase, we first tokenized the generated captions and references using KyTea, and then computed scores.

Hyperparameters We used a gated recurrent unit as recurrent neural network (RNN) cell, and tuned the following hyperparameters: RNN hidden state size, RNN layer size, learning rate, weight decay, dropout probability, beam width, and length normalization coefficient. We chose those with the best CIDEr score on the development set.

Input Modality We used image and motion modalities as encoder inputs. The image modality captures static image content from video frames. In accordance with previous work (Long et al., 2018; Wang et al., 2018), we used the last layer of the ResNet-152 (He et al., 2016) trained on ImageNet.[5] First, we sampled frames at 3 fps and then extracted a 2,048 dimensional vector from each frame. The motion modality captures the local temporal motion. We used 3D ResNeXt-101 (Hara et al., 2018) trained on

[4]https://github.com/google/sentencepiece

[5] https://pytorch.org/docs/master/torchvision/models.html

	PLACE			PERSON			ACTION		
modality	BLEU	ROUGE	CIDEr	BLEU	ROUGE	CIDEr	BLEU	ROUGE	CIDEr
I + M	0.792	0.855	**1.848**	**0.732**	**0.789**	**1.725**	**0.801**	**0.866**	**3.346**
I	**0.833**	**0.868**	1.821	0.717	0.779	1.686	0.780	0.851	3.156
M	0.773	0.830	1.736	0.646	0.736	1.443	0.769	0.844	3.097

Table 3: Results from the phrase generation task. Bold figures indicate the best performer for each evaluation criterion. "I" represents the image modality and "M" is the motion modality.

approach	modality	BLEU	ROUGE	CIDEr
sentence-wise	I + M	0.713	**0.795**	1.837
	I	0.696	0.786	1.769
	M	0.666	0.769	1.677
phrase-wise	I + M	**0.749**	0.791	**1.937**
	I	0.735	0.785	1.846
	M	0.696	0.765	1.729

Table 4: Results from the sentence generation task.

Kinetics-400.[6] We first split a video into a set of 16 frames and then converted each set (16 frames) to a 2,048-dimensional vector. In both modalities, we used mean pooling to aggregate the vectors obtained from a video.

5.2. Experimental Results

Table 3 shows results from the phrase generation task. In all methods except PLACE, the best results were obtained when two modalities were input (I + M). In PLACE, use of the image modality alone was found to be more efficient. This is as expected because generating a phrase for PLACE does not require information about local temporal motion. Consequently, the generator with two modalities did not affect the results.

Table 4 shows the results of sentence generation. The use of two modalities with both sentence-wise and phrase-wise generation performed better than the single modality cases across all criteria, and image modality alone came second.

We found the phrase-wise approach outperformed sentence-wise generations in BLEU and CIDEr. In ROUGE, the sentence-wise approach was observed to be slightly better than the phrase-wise approach.

5.3. Generated Captions

We presented three samples of generated captions and references. Figure 3 shows that the phrase-wise approach captured the action (*blowing a horn*) of the input video, while the sentence-wise approach generated the wrong action phrase (*taking a photo*). Contrary to these results, the captions generated by the sentence-wise approach, shown in figure 4, are better than those of the phrase-wise approach. In Figure 5, neither approach generated accurate action phrases.

6. Conclusion

We constructed a new video caption dataset for describing human actions in Japanese. The advantage of this dataset is that the captions are written in Japanese and specify "who does what and where." To specify this, we conducted two procedures: Phrase annotation and template-based sentence construction. Although the template-based construction does not produce grammatically varied sentences, the sentences produced are not unnatural. Our dataset, consisting of 79,822 videos and 399,233 captions, is the first Japanese caption dataset, and the largest video caption dataset in any language with respect to the number of captions.

We evaluated two approaches based on a multi-modality fusion caption generation method on our dataset: Sentence-wise and phrase-wise approaches. Experiments showed that the phrase-wise approach outperformed the sentence-wise approach with respect to BLEU and CIDEr. In addition, we evaluated phrase generation quality using our dataset, employing phrase generation tasks to ascertain whether the generation methods specified "who does what and where." We observed that the image and motion modalities to be useful in explaining PERSON and ACTION, while image modality alone was sufficient for PLACE.

[6] https://github.com/kenshohara/video-classification-3d-cnn-pytorch

(a) Input video (1 fps).

method	description
sentence-wise	木が生えている屋外で茶色い服を着た男性がカメラで写真を撮っている (A man worn brown clothes is taking a photo by using a camera in the open air with woody)
phrase-wise	屋外で茶色い服の男性が笛を吹いている (A man with brown clothes is blowing a horn in the open air)
Human annotation	屋外でサングラスをした男性が楽器を演奏している (A man worn sunglasses in the open air is playing an instrument)
	外で茶色の服を着ている男性が笛を吹いている (A man is wearing brown clothes who is blowing a horn in the open air)
	屋外で帽子にサングラスをした男性が楽器を演奏している (A man worn a hat and sunglasses is playing an instrument in the open air)
	屋外で帽子とサングラスをした人がしゃがんで角笛を吹いている (A person worn a hat and sunglasses in the open air is squatting eyes and blowing a horn)
	フェンスがある屋外で帽子をかぶってサングラスをかけた男性が楽器を演奏している (A man worn a hat and sunglasses is playing an instrument in the open air with a fence)

(b) Reference descriptions and generated sentences.

Figure 3: An example of ground truth descriptions and sentences generated by the sentence-wise and phrase-wise methods.

(a) Input video (1 fps).

method	description
sentence-wise	車内で坊主頭の男性が食事をしている (A man with a shaven head is eating in a car)
phrase-wise	車内で黒い服を着た男性が話している (A man worn black clothes is speaking in a car)
Human annotation	車内で坊主頭の男性がピザを食べている (A man with a shaven head is eating pizza in a car)
	車内で黒い服の男性が何かを食べている (A man with black clothes is eating something in a car)
	車の中で黒い服を着た男性が何かを食べている (A man worn black clothes is eating something in a car)
	車内で黒い服を着た男性が食べ物を食べている (A man worn black clothes is eating something in a car)
	車内で上着を着た短髪の男性が運転席に座った状態で食べ物を食べている (A short-haired man worn jacket is eating food on the driver's seat in a car)

(b) Reference descriptions and generated sentences.

Figure 4: An example of ground truth descriptions and sentences generated by the sentence-wise and phrase-wise methods.

(c) Input video (1 fps).

method	description
sentence-wise	白い壁の部屋の中で 2 人の男性が抱き合っている (Two men are hugging in a room with white wall)
phrase-wise	白い壁の部屋で迷彩服の男性が抱き合っている (Men with camouflage clothes are hugging in a room with white wall)
Human annotation	屋内で迷彩服の男性が格闘術を教えている (A man with camouflage clothes is teaching hand-to-hand combat inside the room)
	屋内で黒いティーシャツの男性がおさえこまれている (A man with a black T-shirt is being arrested inside the room)
	白い壁の部屋で黒い服を着た男性がサバイバルの訓練をしている (A man worn black clothes is training survival skills in a room with white wall)
	白い壁の部屋の中で黒い服を着た男性が緑色の服を着た男性を殴っている (A man worn black clothes is hitting a man worn green clothes in a room with white walls)
	白い壁で薄暗い屋内でカーキや黒のトップスを着た体格のいい男性たちが護身術を教わっている (Muscular men worn khaki and black clothes are being taught self-defense in a dim room with white wall)

(d) Reference descriptions and generated sentences.

Figure 5: An example of ground truth descriptions and sentences generated by the sentence-wise and phrase-wise methods.

7. Acknowledgements

We thank anonymous reviewers for their valuable comments and suggestions. This paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

8. Bibliographical References

Awad, G., Butt, A., Curtis, K., Lee, Y., Fiscus, J., Godil, A., Joy, D., Delgado, A., Smeaton, A. F., Graham, Y., Kraaij, W., Quénot, G., Magalhaes, J., Semedo, D., and Blasi, S. (2018). TRECVID 2018: Benchmarking Video Activity Detection, Video Captioning and Matching, Video Storytelling Linking and Video Search. In *TRECVID*.

Chen, D. L. and Dolan, W. B. (2011). Collecting Highly Parallel Data for Paraphrase Evaluation. In *ACL*, pages 190–200.

Gan, Z., Gan, C., He, X., Pu, Y., Tran, K., Gao, J., Carin, L., and Deng, L. (2017). Semantic Compositional Networks for Visual Captioning. In *CVPR*, pages 5630–5639.

Gella, S., Lewis, M., and Rohrbach, M. (2018). A Dataset for Telling the Stories of Social Media Videos. In *EMNLP*, pages 968–974.

Hara, K., Kataoka, H., and Satoh, Y. (2018). Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? In *CVPR*, pages 6546–6555.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *CVPR*, pages 770–778.

Jin, Q., Chen, J., Chen, S., Xiong, Y., and Hauptmann, A. (2016). Describing Videos Using Multi-Modal Fusion. In *ACM MM*, pages 1087–1091.

Krishna, R., Hata, K., Ren, F., Fei-Fei, L., and Niebles, J. C. (2017). Dense-Captioning Events in Videos. In *ICCV*, pages 706–715.

Kudo, T. and Richardson, J. (2018). SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In *EMNLP: System Demonstrations*, pages 66–71.

Lin, C.-Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*.

Long, X., Gan, C., and de Melo, G. (2018). Video Captioning with Multi-Faceted Attention. *TACL*, 6:173–184.

Neubig, G., Nakata, Y., and Mori, S. (2011). Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis. In *ACL*, pages 529–533.

Pan, Y., Yao, T., Li, H., and Mei, T. (2017). Video Captioning with Transferred Semantic Attributes. In *CVPR*, pages 984–992.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *ACL*, pages 311–318.

Phan, S., Miyao, Y., and Satoh, S. (2017). MANet: A Modal Attention Network for Describing Videos. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 1889–1894.

Rohrbach, A., Rohrbach, M., Qiu, W., Friedrich, A., Pinkal, M., and Schiele, B. (2014). Coherent Multi-sentence Video Description with Variable Level of Detail. In *German Conference on Pattern Recognition*, pages 184–195.

Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A., and Schiele, B. (2017). Movie Description. *IJCV*, 123(1):94–120.

Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I., and Gupta, A. (2016). Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In *ECCV*, pages 510–526.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). CIDEr: Consensus-Based Image Description Evaluation. In *CVPR*, pages 4566–4575.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., and Saenko, K. (2015a). Sequence to Sequence-Video to Text. In *ICCV*, pages 4534–4542.

Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., and Saenko, K. (2015b). Translating Videos to Natural Language Using Deep Recurrent Neural Networks. In *NAACL*, pages 1494–1504.

Wang, X., Wang, Y.-F., and Wang, W. Y. (2018). Watch, Listen, and Describe: Globally and Locally Aligned Cross-Modal Attentions for Video Captioning. In *NAACL*, pages 795–801.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation. *arXiv preprint arXiv:1609.08144*.

Xu, J., Mei, T., Yao, T., and Rui, Y. (2016). MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *CVPR*, pages 5288–5296.

Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., and Courville, A. (2015). Describing Videos by Exploiting Temporal Structure. In *ICCV*, pages 4507–4515.

Yoshikawa, Y., Lin, J., and Takeuchi, A. (2018). STAIR Actions: A Video Dataset of Everyday Home Actions. *arXiv preprint arXiv:1804.04326*.

Zhou, L., Xu, C., and Corso, J. J. (2018). Towards Automatic Learning of Procedures From Web Instructional Videos. In *AAAI*, pages 7590–7598.