Fine-Grained Action Understanding with Tools in Instructional Videos

Saelyne Yang¹  Jaesang Yu¹  Jae Won Cho²  Juho Kim¹

¹KAIST  ²Sejong University

Abstract

Instructional videos provide step-by-step information on how to achieve a task. Previous research in video understanding has advanced comprehension of the hierarchical structure of procedural information in videos, such as goals and steps, by introducing several datasets and models. While the emphasis has largely been on the ‘what’ aspects of actions (e.g., frying eggs, cutting a carrot), it is also important to understand the ‘how’ behind these actions. In this research, we focus on tools (e.g., comb, hair clip) used to perform tasks for a fine-grained understanding of actions, as tools have direct relevance to how actions are performed and drive the changes of the final outcome of the task. To achieve this, we developed an annotation schema that identifies the tools used in a step and how the tool is used for which purpose. Based on the schema, we annotated 48 video clips across 12 domains, each corresponding to a step in COIN [23], an instructional video dataset with taxonomy-based step labels. Our collected dataset reveals the detailed granularity of actions, demonstrating diversity even within identical steps. Through an analysis of the dataset, we demonstrate the significance of tool-centered annotation for fine-grained action understanding.

1. Introduction

Instructional videos provide visual demonstrations of how to achieve tasks, such as ‘Cooking pasta’, often complemented by verbal instructions. These videos provide step-by-step guidance toward achieving task goals, containing hierarchical and procedural knowledge. To facilitate procedural video learning, various datasets have been introduced [10, 14, 21, 23, 26, 27, 32, 34, 35]. These datasets are annotated with temporal segment boundaries and the actions performed within each segment, enabling a range of video understanding tasks such as video or moment retrieval [2, 12, 14, 29], video captioning [1, 9, 13, 19, 25, 30], and action recognition or localization [3–5, 8, 11, 16, 18, 24].

While these datasets have advanced video understanding, they primarily focus on the ‘what’ aspects of actions rather than the ‘how’. Comprehending how actions are performed is vital for understanding the actions as this lies at the core of skill acquisition and human intelligence [20, 22, 28]. Improved perception of the nuances in actions would facilitate both humans and robots to replicate these actions better [6]. Recent efforts in fine-grained action understanding have incorporated learning verb-adverb relationships to distinguish between actions such as ‘slice slowly’ and ‘slice quickly’ [6, 7, 15]. While they focus on elaborating the action through the use of adverbs, there are much more detailed aspects of how an action is performed. For instance, an action like ‘Cutting a carrot’ could entail details such as the tool used or the shape of the cut.

In this research, we propose that incorporating tools—the equipment used in task execution—can serve as a means of understanding how actions are performed in more detail. We focus on instructional videos that involve physical demonstrations such as cooking and crafting, which often employ tools to accomplish tasks [31]. The use of tools, whether directly or indirectly, results in changes to the target object of the task, thereby enabling us to understand the step in actions that drive these changes. For example, understanding the action ‘Cutting a carrot’ in terms of the tools used would enable us to discern the specific tools employed (such as knives or scissors) and how they are utilized.

To enhance the fine-grained understanding of instructional videos with tools, we introduce a tool-oriented annotation schema and curate a dataset containing action-level descriptions. Based on the COIN dataset [23], an instructional video caption dataset with task and step-level annotations, as a source of videos and captions, we choose two of the step labels from distinct tasks from each of the 12 domains and pair each step with two distinct video clips. With our annotation schema, we produce tool-oriented descriptions for each step, beginning with identifying tools used, specifying the actions performed with each tool, and elaborating on how each tool is used, resulting in a free-form text annotation within a syntactic structure.

Ultimately, we collected 186 action-level descriptions, leading to a total of 1.7K words¹. There is an average of 3.9

¹https://github.com/saelyne/how2how2
actions per step label and an average of 9 words within each annotated action description. Through extensive analysis, we demonstrate the significance of tool-centered annotation for fine-grained procedural video understanding.

2. Dataset Collection

We constructed a dataset that contains action information based on tools used in instructional videos. Below, we describe how we selected the videos and discuss our tool-oriented annotation schema.

2.1. Video Selection

Existing datasets on instructional videos have been annotated using predefined steps [23, 35] or via free-from text [32, 34]. We chose to start with a set of videos annotated with predefined steps and further elaborate on the steps with free-form text annotations, thereby creating a hierarchical structure while ensuring diverse and high-quality captions. We selected videos from the COIN dataset [23], given its coverage of various domains and its hierarchical organization of tasks (e.g., ‘Replace a Bulb’) and steps (e.g., ‘Take out the old bulb’) in each task.

We first randomly sampled two tasks from each of the 12 domains in COIN. For each task, we select one of the predefined steps. Here, we avoid selecting steps that are too broad, such as ‘Make the detergent’. To further analyze the differences in a step with more granularity, we chose two video clips for each step. This resulted in a collection of 48 video clips, spanning a diverse range of tasks and domains. Table 1 provides details regarding the step labels we select.

2.2. Tool-oriented Annotation Schema

For fine-grained action annotation, we focus on the tools used and their role in each step. Since the annotation framework significantly impacts response quality, each component should be designed with substantial consideration [17]. In this work, as a first step, we as the authors have taken on the role of annotators and iteratively revised its design. As a result, we have developed a cascaded annotation schema comprising three stages, with each stage building upon the previous one. Our goal is to keep each step straightforward, enabling annotators to concentrate and build upon each response, and therefore induce enriched annotations.

Figure 1 shows the overall annotation schema we devised, which focuses on annotating fine-grained actions with tool information. Starting from Stage 1, annotators are asked to identify the tools used in the step given a video clip. Here, tools are defined as objects directly used by the demonstrator to execute the step, while materials that are integrated to produce the final target object, such as ingredients, are not considered as tools. Then in Stage 2, annotators are prompted to specify what the demonstrator does with each tool (Tool Usage action). Note that the same tool might be used in different actions, and an action might require the use of multiple tools. Stage 3 asks annotators to describe how the demonstrator uses the tool, thereby elaborating the tool usage (Using Tool action). Here, we ask them to describe these aspects from a first-person perspective to ensure consistency within the descriptions. Ultimately, each elaborated action description is formed by combining the Tool Usage action and the Using Tool action alongside the
specified tool. Our schema leverage the advantages of both fixed-set \cite{23, 35} and free-form annotations \cite{32, 34}, which allow for efficient learning for various video understanding tasks while ensuring elaboration of descriptions. After finalizing each description, annotators are asked to provide the corresponding time segment.

3. Dataset Analysis

With the selected steps and annotation schema, we have gathered a total of 186 action-level descriptions, amounting to 1.7K words, as a preliminary result reflecting our initial steps. On average, each step label contains 3.9 actions, with each action description comprising 9 words. Our annotation schema, designed in a cascaded fashion, facilitated annotation with a broad vocabulary. Furthermore, the schema allows free-form natural language descriptions, resulting in a dataset consisting of 128 unique sentences and a vocabulary of 228 words. Calculating the ratio of the number of sentences to the size of vocabulary as a measure of sentence diversity \cite{33}, we found that YouCook2 \cite{34} and HiREST \cite{32}, which contain action-level descriptions, have averages of 5.4 and 2.3 sentences per word, respectively. Whereas our dataset exhibits an average of 0.8 sentences per word, indicating a high degree of sentence diversity compared to previous instructional video datasets.

Our annotation schema emphasizes the elaboration of actions, focusing on both the ‘what’ and ‘how’ aspects of tool usage. These are categorized as Tool Usage actions, describing what the person does with the tool, and Using Tool actions, explaining how the person employs the tool for its intended purpose. As a result, each action description in our dataset follows a structure, comprising three parts: the tool itself, a description of its usage, and a description of how it is used. This breakdown enables a comprehensive analysis of how tools are integrated into task demonstrations within an instructional context.

Figure 2 shows how the first three words of our action descriptions vary, indicating that tool-oriented descriptions have a diverse vocabulary within a fixed syntactic framework. Specifically, the Tool Usage verbs (Fig. 2a), which detail the purpose of the tools, convey high-level actions and the intentions behind their use in achieving the conceptual goal (e.g., remove, distribute, straighten, etc.). Conversely, the Using Tool verbs (Fig. 2b), which illustrate the method of using the tools, are associated with lower-level actions, focusing on the fundamental atomic actions undertaken with the tool (e.g., pulling, pushing, pressing, etc.).

It is also important to understand both what the tools are used for and how they are used. For a given tool, there may be various methods of utilization. To examine the diversity in how each tool is employed, we have analyzed three components for each action-level annotation: the tool used, the verb describing the tool’s purpose (Tool Usage verb), and the verb detailing how the tool is used (Using Tool verb). These components were combined to create a unique set, and we counted the occurrences of each set across all tasks and domains. Each set was treated as unique for each video.

Figure 2. Action descriptions in our dataset can be decomposed into three components: (1) Tool, (2) Tool Usage (what the demonstrator does with the tool), and (3) Using Tool (how the person uses the tool). The figures describe distributions of natural language descriptions for Tool Usage and Using Tool actions by their first three words. We removed stop words and stemmed the remaining words. The stemmed words appearing multiple times in one video were counted as unique to analyze the general distribution of the dataset without any bias.
clip to focus on the general distribution. Figure 3 illustrates the overall distribution in the gathered set. This visualization demonstrates that the same tool can be utilized for different purposes (Tool Usage), and even when tools are employed for the same purpose, there can be different methods of usage (Using Tool).

4. Discussion

As we conducted annotation for two distinct video clips corresponding to the same step label within the same task, we were able to analyze how our annotation schema can effectively distinguish between them by breaking down the step into tool-oriented actions. In our dataset, for instance, consider the step label ‘Open the bottle carefully’, within the ‘Open Champagne Bottle’ task. One video clip depicts a bottle being opened using a knife, with the demonstrator vigorously scratching the side of it back and forth, while the other video clip shows a bottle being opened using hands, with the demonstrator gently twisting the bottom part of the bottle. Since each video employs different tools to perform the task, this results in a completely different set of actions. Therefore, we argue that capturing the tools used to perform a task or action is a crucial aspect in designing an instructional video dataset for fine-grained action understanding.

Another issue worth discussing is that our annotation schema has been designed to adhere to a fixed syntactic structure for annotations. The action descriptions in our dataset mostly follow a consistent grammar format: Tool Usage action (verb + noun) + with + Tool (noun) + by + Using Tool action (verb + noun + adverb) (e.g., ‘clean the floor with a brush by distributing the solution outwardly’), except a few samples that also include prepositional information (e.g., “put the string in between the tuning post with hands by pressing it evenly”), and some without a description of how the tool is used (Using Tool). While we have observed that some step-level descriptions are indivisible, resulting in a single action description, this single action description serves to elaborate upon the original step label. For example, the original step label “push curling” has been expressed as a single action description “push the curling stone with hands by giving a clockwise spin”. Thus, our tool-oriented schema was able to facilitate rich elaborations at all levels of step complexity in the COIN dataset.

5. Conclusion and Future Directions

Our tool-oriented elaborated annotations offer valuable resources for advancing language-based procedural video understanding. The atomic action-level descriptions, enriched with a diverse vocabulary from our proposed annotation schema, have the potential to enhance video understanding in various downstream tasks, such as dense video captioning, action localization, and multi-label action recognition. Additionally, our tool-oriented dataset opens up opportunities for novel tasks in tool learning for instructional tasks. Given the crucial role of tools in instructional actions, as highlighted by our analysis, we plan to propose tasks such as tool usage recognition and learning, further enhancing the use of procedural videos.

Moving forward, we are in the process of curating a comprehensive instructional video dataset utilizing our tool-oriented annotation framework. This dataset will cover various domains within instructional videos and include videos that feature a wide range of tools. To demonstrate the effectiveness of our curated dataset and proposed annotation schema, we intend to design a hierarchical benchmark comprising multi-level descriptions. Additionally, we aim to introduce a novel video-tool learning task, which will pave the way for a deeper understanding of procedural tasks through videos.

Acknowledgements This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2021-0-01347, Video Interaction Technologies Using Object-Oriented Video Modeling).
References


[9] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan CarlosNiebles. Dense-captioning events in videos. In ICCV, 2017. 1

[10] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan CarlosNiebles. Dense-captioning events in videos. In ICCV, 2017. 1


[16] Ioanna Ntinou, Enrique Sanchez, and Georgios Tzimiropoulos. Multiscale vision transformers meet bipartite matching for efficient single-stage action localization. In CVPR, 2024. 1


[27] Yi Wang, Yinan He, Yizhuo Li, Yizhuo Li, Kunchang Li, Jiaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Intervid: A large-scale video-text dataset for multimodal understanding and generation. ICLR, 2024. 1


[29] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large-scale dataset for video-subtitle moment retrieval. In ICCV, 2017. 1


[34] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In AAAI Conference on Artificial Intelligence, pages 7590–7598, 2018. 1, 2, 3

6. Selected Step Information
<table>
<thead>
<tr>
<th>Domain</th>
<th>Task</th>
<th>Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nursing and Care</td>
<td>PutOnHairExtensions</td>
<td>pull up the hair to reserve place for the hair extensions</td>
</tr>
<tr>
<td></td>
<td>RemoveBlackheadsWithGlue</td>
<td>wipe the glue to a layer</td>
</tr>
<tr>
<td>Sport</td>
<td>PlayCurling</td>
<td>push curling</td>
</tr>
<tr>
<td></td>
<td>ThrowHammer</td>
<td>pre-swing</td>
</tr>
<tr>
<td>Vehicle</td>
<td>InstallBicycleRack</td>
<td>mount the bracket to the back of the car</td>
</tr>
<tr>
<td></td>
<td>ReplaceAWiperHead</td>
<td>take out the wiper</td>
</tr>
<tr>
<td>Science and Craft</td>
<td>MakeSlimeWithGlue</td>
<td>rub and drag the materials</td>
</tr>
<tr>
<td></td>
<td>MakeFlowerCrown</td>
<td>stick or bind flower to the frame</td>
</tr>
<tr>
<td>Electrical Appliance</td>
<td>CleanLaptopKeyboard</td>
<td>clean the inside of the button</td>
</tr>
<tr>
<td></td>
<td>MakeRJ45Cable</td>
<td>cut a certain length</td>
</tr>
<tr>
<td>Furniture and Decoration</td>
<td>InstallClosestool</td>
<td>connect the water pipe</td>
</tr>
<tr>
<td></td>
<td>ReplaceDoorKnob</td>
<td>install the new door knob</td>
</tr>
<tr>
<td>Gadgets</td>
<td>MakeWirelessEarbuds</td>
<td>process the copper wire inside the earphone cable</td>
</tr>
<tr>
<td></td>
<td>OpenALockWithPaperclips</td>
<td>twist the paperclips by hands</td>
</tr>
<tr>
<td>Drink and Snack</td>
<td>MakeHomemadeIceCream</td>
<td>stir the mixture</td>
</tr>
<tr>
<td></td>
<td>MakeStrawberrySmoothie</td>
<td>put strawberries and other fruits into the juicer</td>
</tr>
<tr>
<td>Leisure and Performance</td>
<td>ChangeGuitarStrings</td>
<td>fix the new string on the head of the guitar</td>
</tr>
<tr>
<td></td>
<td>OpenChampagneBottle</td>
<td>open the bottle carefully</td>
</tr>
<tr>
<td>Housework</td>
<td>IronClothes</td>
<td>iron the cloths with the iron</td>
</tr>
<tr>
<td></td>
<td>CleanCementFloor</td>
<td>clean the floor</td>
</tr>
<tr>
<td>Dish</td>
<td>CookOmelet</td>
<td>fry eggs</td>
</tr>
<tr>
<td></td>
<td>UseRiceCookerToCookRice</td>
<td>soak and wash the rice</td>
</tr>
<tr>
<td>Pets and Fruit</td>
<td>Transplant</td>
<td>take out the plant</td>
</tr>
<tr>
<td></td>
<td>Sow</td>
<td>sow on the soil</td>
</tr>
</tbody>
</table>

Table 1. Selected task and step labels within each domain from the COIN dataset