Enhancing How People Learn Procedural Tasks Through How-to Videos

Saelyne Yang
School of Computing, KAIST
Daejeon, South Korea
saelyne@kaist.ac.kr

ABSTRACT

Humans learn skills to perform various tasks in their everyday lives. While how-to videos serve as a popular tool for people to learn skills and achieve tasks, there are limitations in learning from videos such as difficulties in accessing information in need or lack of personalized support. My Ph.D. research aims to enhance how people learn procedural tasks through how-to videos by understanding and improving the consumption of video content, application of the content to their own context, and reflection on the experiences. This research presents opportunities and insights into how we could better leverage videos for humans to learn skills and achieve tasks.

KEYWORDS

procedural task, how-to videos, video interaction, skill learning

ACM Reference Format:
Saelyne Yang. 2024. Enhancing How People Learn Procedural Tasks Through How-to Videos. In *The 37th Annual ACM Symposium on User Interface Software and Technology (UIST Adjunct '24), October 13–16, 2024, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3672539.3686711

1 INTRODUCTION

Skill acquisition is essential in our lives. It enables us to achieve our goals, from completing everyday tasks to fulfilling our dreams. People learn various skills, such as cooking a pasta dish or playing the piano, to enhance their capabilities and pursue their passions.

Videos have become an invaluable resource for learning new skills [1]. "How-to" videos, in particular, provide step-by-step instructions for procedural tasks such as cooking, makeup application, and crafting. These videos offer detailed visual workflows and verbal explanations, helping learners understand and follow the tasks.

However, there are several limitations to learning from videos. First, it is difficult to skim through and navigate them, making it hard for users to grasp the overall content and find specific points of interest. Second, because videos are created for a broad audience, it is challenging to receive personalized support. Additionally, due to the asynchronous nature of video-on-demand, getting real-time help when users encounter difficulties in understanding or following the content can be difficult.

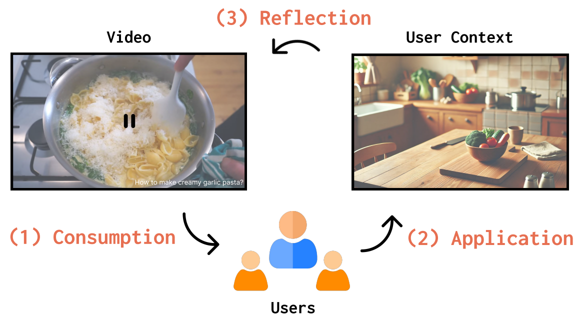


Figure 1: My research focuses on enhancing how people learn procedural tasks through how-to videos, by understanding and improving the lifecycle of video learning: consumption, application, and reflection.

My Ph.D. research aims to address these limitations and enhance how people learn procedural tasks through how-to videos. To achieve this, my research spans the lifecycle of video learning (Figure 1): **(1) Consumption**: users first watch and consume the video content to understand the instructions and information being presented. **(2) Application**: users then apply the instructions from the video to their own context, replicating or adapting the demonstrated tasks. **(3) Reflection**: finally, users reflect on the experience, evaluate how it went in their own context, and formulate questions or seek further clarification. My research explores each of these phases and develops systems to improve them, contributing to a more effective video-based learning experience.

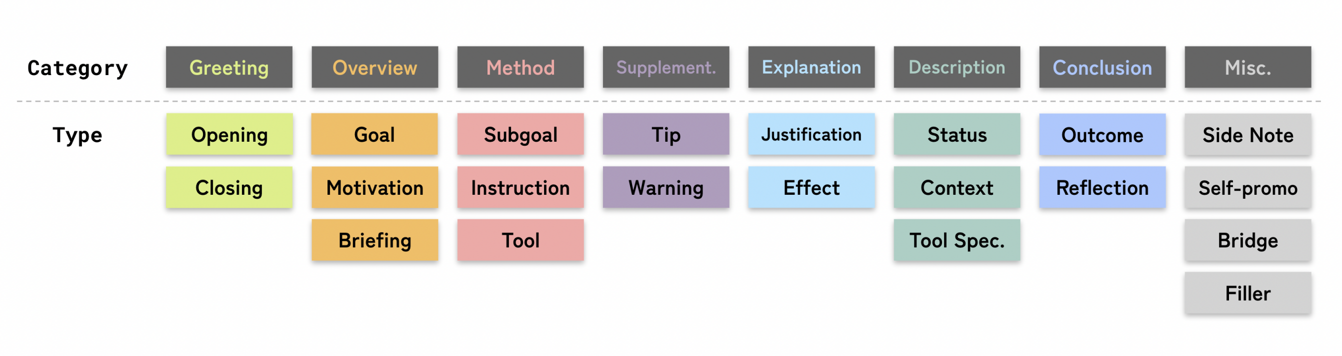
In light of the emergence of AI systems and agents that assist users in performing tasks, my research explores how we can better use video as a learning resource in performing tasks. It provides insights into maximizing the use of videos and enhancing their role in skill acquisition, thereby contributing to the broader study of the role of computers and AI in human skill development.

2 VIDEO CONTENT CONSUMPTION

The first phase in the video learning cycle is to consume the video content. In other words, users first watch the video to understand the instructions and information being presented. However, how-to videos are rich in information—they not only give instructions, but also provide justifications for a particular action or descriptions of tools they use. People might seek different information to meet their needs, but the diverse information within a video is scattered throughout, making it difficult for users to identify the information in need.

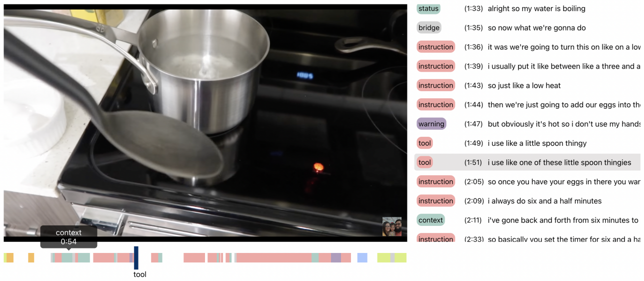
UIST Adjunct '24, October 13–16, 2024, Pittsburgh, PA, USA

ACM ISBN 979-8-4007-0718-6/24/10
https://doi.org/10.1145/3672539.3686711



Category	Greeting	Overview	Method	Supplement.	Explanation	Description	Conclusion	Misc.
Type	Opening	Goal	Subgoal	Tip	Justification	Status	Outcome	Side Note
	Closing	Motivation	Instruction	Warning	Effect	Context	Reflection	Self-promo
		Briefing	Tool			Tool Spec.		Bridge
								Filler

Figure 2: Taxonomy of information types in how-to videos [5].



status (1:33) alright so my water is boiling
bridge (1:35) so now what we're gonna do
instruction (1:36) it was we're going to turn this on like on a lov
instruction (1:39) i usually put it like between like a three and a
instruction (1:43) so just like a low heat
instruction (1:44) then we're just going to add our eggs into th
warning (1:47) but obviously it's hot so i don't use my hands
tool (1:49) i use like a little spoon thingy
tool (1:51) i use like one of these little spoon thingies
instruction (2:05) so once you have your eggs in there you war
instruction (2:09) i always do six and a half minutes
context (2:11) i've gone back and forth from six minutes to
instruction (2:33) so basically vou set the timer for six and a ha

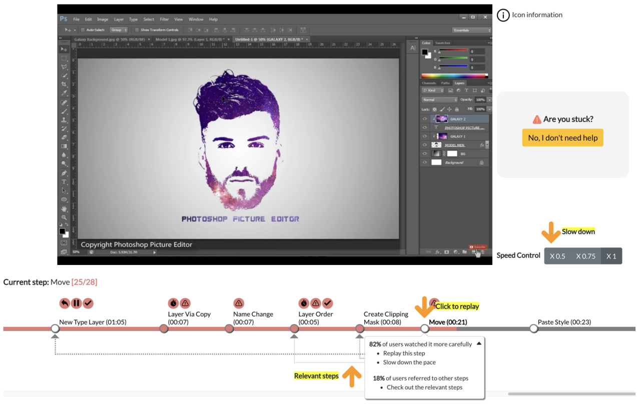
Figure 3: A video interface that supports navigation based on information types [5].

I proposed that a comprehensive taxonomy that identifies and categorizes the types of information shared in how-to videos can serve as a foundation for supporting users in navigating videos. Thus, I presented a taxonomy of information types in how-to videos [5] (Figure 2). To construct the taxonomy, I selected 120 videos from the HowTo100M dataset [3] and performed an iterative open coding of 4k sentences from 48 videos. From the analysis, 21 information types emerged under 8 categories: *Greeting, Overview, Method, Supplementary, Explanation, Description, Conclusion, and Miscellaneous.* The utility of the taxonomy was demonstrated in both analyzing users' navigational behavior and supporting their navigation in how-to videos. I observed that existing video systems built to support navigation utilized different information types to meet users' specific needs. Furthermore, I built a research probe that enables users to navigate using the information types within the video (Figure 3). A user study showed that the participants effectively used different information types for finding specific information needed to perform each of the Search, Summarize, and Follow tasks. This taxonomy enables a wide range of video-related tasks beyond navigation, such as video authoring, viewing, and analysis.

3 APPLICATION TO USER CONTEXT

After obtaining the content from videos, users apply it to their own context. For example, in software tutorials such as Photoshop tutorials, users watch the video for instructions and then apply them to the target software. However, this process often requires frequent back-and-forth between the video and the application, which incurs cognitive overhead. Additionally, users need to constantly compare the video instructions with their own work to ensure they are following them correctly, as they are prone to missing out on subtle differences.

To address this problem, I proposed SoftVideo [8] (Figure 4), a system that helps users plan ahead before watching each step in tutorial videos and provides feedback and help to users on their progress. SoftVideo is powered by collective interaction data, as experiences of previous learners with the same goal can provide insights into how they learned from the tutorial. By identifying the difficulty and relatedness of each step from the interaction logs, SoftVideo provides information on each step such as its estimated difficulty, lets users know if they completed or missed a step, and suggests tips such as relevant steps when it detects users struggling. To enable such a data-driven system, I collected and analyzed video interaction logs and the associated Photoshop usage logs for two tutorial videos from 120 users. I then defined six metrics that portray the difficulty of each step, including the time taken to complete a



Are you stuck?
No, I don't need help

Current step: Move [25/28]

82% of users watched it more carefully
- Replay this step
- Slow down the pace

18% of users referred to other steps
- Check out the relevant steps

Figure 4: SoftVideo [8] provides step information, gives feedback to learners on their progress, and provides help to overcome confusing moments by tracking the progress in their video and their own software.

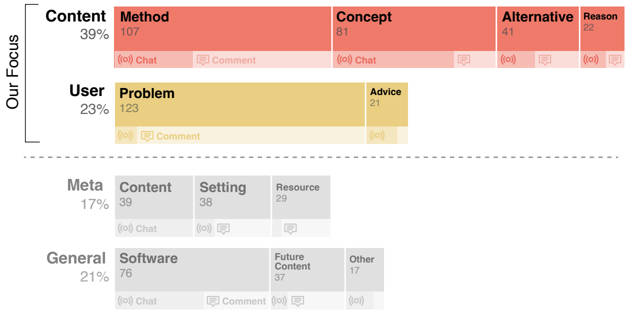
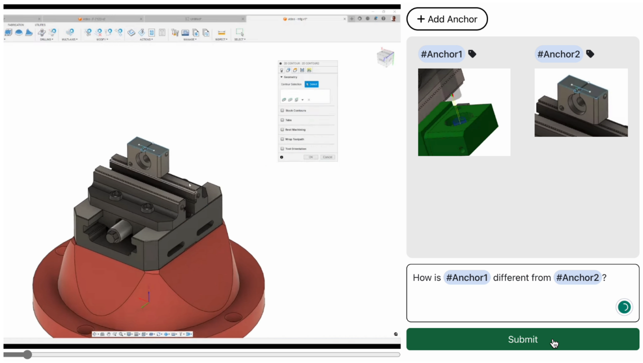


Figure 5: Categories and types of questions identified from an analysis of live chat and comment data of software tutorial videos [7].



How is #Anchor1 different from #Anchor2 ?

Figure 6: A system where users can ask questions by directly referring to specific (visual) parts of the video [7].

step and the number of pauses in a step, which were also used to detect users’ struggling moments by comparing their progress to the collected data. A user study showed that that participants could proactively and effectively plan their pauses and playback speed, and adjust their concentration level. They were also able to identify and recover from errors with the help SoftVideo provides.

4 REFLECTION ON THE CONTENT

Lastly, after users watch a video and apply the content to their own context, they may reflect on the material and formulate questions or seek further clarification. I studied the types of questions users ask, the ways in which they ask them, and how to design a system that provides answers. I describe two of my projects below.

4.1 Question-Answering with Visual Anchors

While tutorial videos are a popular help source for learning feature-rich software, getting quick answers to questions about tutorial videos is difficult. I presented an automated approach for responding to tutorial questions [7]. I first identified different question types by analyzing 633 questions found in 5,944 video comments (Figure 5). Focusing on types that are relevant to the video content, I observed that users frequently described parts of the video in questions. To further delve into the types of visual content that users reference in their questions, I developed a system that allows users to ask questions by creating visual anchors (Figure 6). Through the system, 217 questions were collected, each accompanied by one or more visual anchors. Based on the finding that most visual anchors referred to UI elements and the application workspace, I built AQuA, a pipeline that generates useful answers to questions with visual anchors. I demonstrated this for Fusion 360, showing that we can recognize UI elements in visual anchors and generate answers using GPT-4 augmented with that visual information and software documentation. An evaluation study demonstrated that this approach provides better answers than baseline methods.

4.2 Video Question Answerability and Roles of Modalities in Answers

A number of computational models have been developed for Video Question Answering (Video QA) tasks in the ML and CV communities, aiming to provide immediate answers to questions users have about the video. However, they are primarily trained on questions that are generated *from* the video content, producing answers from within the content. However, in real-world situations, users may pose questions that go beyond the video’s informational boundaries, highlighting the necessity to determine if a video can provide the answer. I presented YTCommentQA [6], a dataset that contains naturally generated questions from YouTube, with an indication of whether the question is answerable within the video or not. It also categorizes answerable questions based on the required modality to answer — visual, script, or both. The analysis demonstrated that some information is complemented by both visual and script elements (Figure 8). Experiments with answerability classification

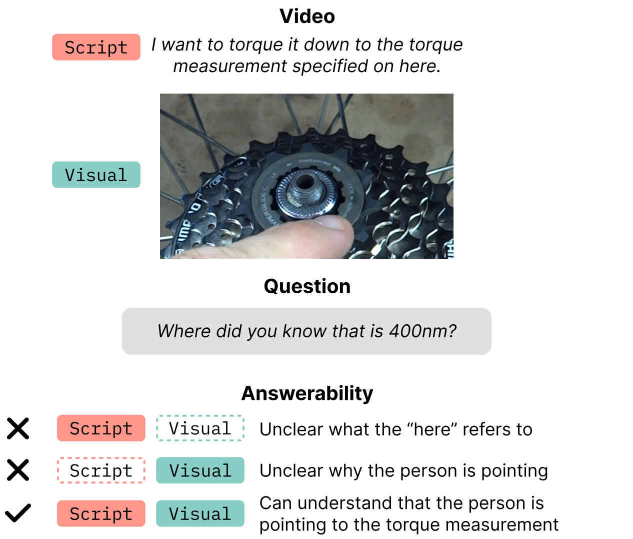

Video
Script
I want to torque it down to the torque measurement specified on here.
Visual
Question
Where did you know that is 400nm?
Answerability
✗ Script Visual Unclear what the “here” refers to
✗ Script Visual Unclear why the person is pointing
✓ Script Visual Can understand that the person is pointing to the torque measurement

Figure 7: A question on video can be either (1) unanswerable by video, (2) answerable by visual, (3) answerable by script, or (4) answerable when both visual and script are present. The figure shows an example of (4), where the question is answerable with the understanding of both visual and script [6].

tasks highlight the complexity of YTCommentQA, emphasizing the need to comprehend the combined role of visual and script information in video reasoning. This work provides the HCI perspective on approaching Video QA tasks.

5 CURRENT AND FUTURE DIRECTIONS

In this section, I outline the research topics I am currently exploring and discuss future research agendas.

5.1 Capturing Tacit Knowledge in Videos

What makes people truly learn a skill? There are many instances where users follow video instructions but fail to perform the task successfully or do not achieve the same quality as demonstrated. I believe tacit knowledge, which is implicitly shared or demonstrated rather than explicitly taught, plays a crucial role. For example, in the instruction "*sprinkle flour over potatoes*", tacit knowledge might involve the subtle technique of tapping the flour tray while sprinkling to ensure even distribution. This type of knowledge is essential, often influencing the quality of the outcome. Then, how can we capture such tacit knowledge? My work-in-progress paper [9] has developed a schema for fine-grained action understanding. By analyzing the detailed actions and their effects, we can capture the subtle actions that are meaningful in achieving the task. I am currently extending this line of work, believing that understanding the nuances of a task can significantly help users perform the task.

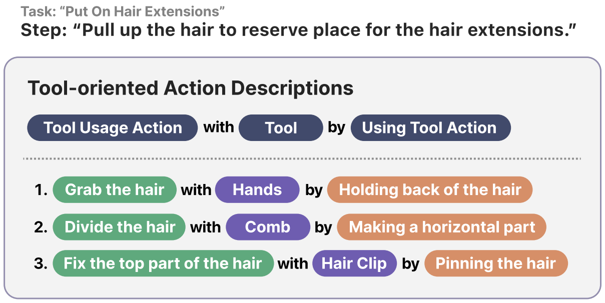


Figure 8: A tool-oriented annotation schema designed to capture fine-grained actions in how-to videos.

5.2 Providing Adaptive Video Content Based on User Context

Another important aspect of skill learning is that the learning materials should match the user's context. This context includes anything related to the user's setting where they apply the task, such as available tools or prior knowledge of the task. To support this, I have conducted a study observing how users navigate through multiple videos on a task to learn and align the material to their needs. Based on these insights, I am designing a system that provides adaptive video content by synthesizing content from multiple video sources. I am actively exploring how we can provide more personalized learning experiences through videos, akin to how large language models generate customized responses to user inputs.

6 BROADER IMPACT

In this section, I discuss the broader impact that my research brings in the video creation and skill learning with VR/AR techniques.

6.1 Beyond Consumption: Towards Effective Video Creation

Recent advances in Generative AI have enabled the creation of videos ranging from entertaining to instructional content. My research provides insights into what makes videos effective for learning and can guide the video generation process. For instance, video generation models can be designed to produce both visual and verbal information, and my work on information types [5] helps determine which information should be highlighted during this process. These principles can be applied not only to video generation but also to video editing. I have explored methods to make video editing more accessible [2] and effective using natural language and sketching modalities [4]. With these techniques, we can develop a more accessible and user-friendly video editing system that suggests edits to enhance the video content or even generates multiple versions of videos tailored to users' preferences and needs.

6.2 Beyond Videos: Towards Comprehensive Support for Human Skill Learning

Recent advances in AI have spurred significant research on supporting humans performing tasks in various forms, such as through VR/AR. I envision that integrating video resources with these external tools and interfaces can offer more comprehensive support for human skill learning. For example, users with smart glasses can receive direct assistance while performing tasks, with the AI suggesting relevant video segments for the current step or providing help to overcome mistakes. With the integration of other modalities, I believe that the use of videos will increase significantly, offering a richer and more effective learning experience.

7 CONCLUSION

In this paper, I present my research for enhancing learners' ability to perform procedural tasks through how-to videos. I described my past research that seeks to achieve this goal through understanding and improving the lifecycle of video learning, from consumption to application and reflection. I also outlined my current and future directions to support skill learning more comprehensively. Through this research, I hope to enable more people to learn skills better and complete everyday tasks.

ACKNOWLEDGMENTS

I am grateful to my advisor, Juho Kim, for his continued support and guidance in my research. I also want to thank my collaborators in Adobe, Autodesk, and LG AI research, as well as the KIXLAB members for their support and feedback. My research was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No.2021-0-01347, Video Interaction Technologies Using Object-Oriented Video Modeling).

REFERENCES

[1] Pei-Yu Chi, Sally Ahn, Amanda Ren, Mira Dontcheva, Wilmot Li, and Björn Hartmann. 2012. MixT: automatic generation of step-by-step mixed media tutorials. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology* (Cambridge, Massachusetts, USA) *(UIST '12)*. Association for Computing Machinery, New York, NY, USA, 93–102. https://doi.org/10.1145/2380116.2380130

[2] Mina Huh, Saelyne Yang, Yi-Hao Peng, Xiang 'Anthony' Chen, Young-Ho Kim, and Amy Pavel. 2023. AVscript: Accessible Video Editing with Audio-Visual Scripts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 796, 17 pages. https://doi.org/10.1145/3544548.3581494

[3] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.

[4] Bekzat Tilekbay, Saelyne Yang, Michal Adam Lewkowicz, Alex Suryapranata, and Juho Kim. 2024. ExpressEdit: Video Editing with Natural Language and Sketching. In *Proceedings of the 29th International Conference on Intelligent User Interfaces (IUI '24)*. ACM. https://doi.org/10.1145/3640543.3645164

[5] Saelyne Yang, Sangkyung Kwak, Juhoon Lee, and Juho Kim. 2023. Beyond Instructions: A Taxonomy of Information Types in How-to Videos. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 797, 21 pages. https://doi.org/10.1145/3544548.3581126

[6] Saelyne Yang, Sunghyun Park, Yunseok Jang, and Moontae Lee. 2024. YTCommentQA: Video Question Answerability in Instructional Videos. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 17 (Mar. 2024), 19359–19367. https://doi.org/10.1609/aaai.v38i17.29906

[7] Saelyne Yang, Jo Vermeulen, George Fitzmaurice, and Justin Matejka. 2024. AQuA: Automated Question-Answering in Software Tutorial Videos with Visual Anchors. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 928, 19 pages. https://doi.org/10.1145/3613904.3642752

[8] Saelyne Yang, Jisu Yim, Aitolkyn Baigutanova, Seoyoung Kim, Minsuk Chang, and Juho Kim. 2022. SoftVideo: Improving the Learning Experience of Software Tutorial Videos with Collective Interaction Data. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 646–660. https://doi.org/10.1145/3490099.3511106

[9] Saelyne Yang, Jaesang Yu, Jae Won Cho, and Juho Kim. 2024. Fine-Grained Action Understanding with Tools in Instructional Videos. In *CVPR 2024 Workshop on Learning from Procedural Videos and Language*.