# Improving Video Interfaces by Presenting Informational Units of Videos

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

Sangkyung Kwak
skkwak9806@kaist.ac.kr
School of Computing, KAIST
Daejeon, Republic of Korea

Tae Soo Kim
taesoo.kim@kaist.ac.kr
School of Computing, KAIST
Daejeon, Republic of Korea

Juho Kim
juhokim@kaist.ac.kr
School of Computing, KAIST
Daejeon, Republic of Korea

## ABSTRACT

Videos have been a major resource that people use when they seek information. How-to videos or instructional tutorials provide detailed visual explanations to convey procedural knowledge. However, existing video interfaces that only contain a linear timeline make it difficult for users to navigate the content of the video and identify meaningful segments. To overcome such limitations, previous work has shown that identifying *informational units* of videos can help users navigate the video content. In this work, we investigate the effect of several informational units in the video: objects, actions, and relations between actions. Through three formative studies, we present how providing such information in video interfaces changes users' experiences, and what challenges may arise from having this information. Finally, we discuss how our findings align with human event perception theory and possible directions for future video interfaces.

## CCS CONCEPTS

• **Human-centered computing → Interactive systems and tools**.

## KEYWORDS

video navigation, video interaction

**ACM Reference Format:**
Saelyne Yang, Sangkyung Kwak, Tae Soo Kim, and Juho Kim. 2022. Improving Video Interfaces by Presenting Informational Units of Videos. In *Proceedings of CHI Conference on Human Factors in Computing Systems (CHI '22 Extended Abstracts).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

Recently, people seek and consume information through videos, instead of traditional text-based materials. This has especially been the case for procedural information such as how-tos or instructional tutorials [10]. While videos as an interface are effective at demonstrating relevant visual and temporal details, they have a major flaw: information is encoded as a linear and continuous stream. While most GUIs are segmented into sections, components, and pages, videos are an array of continuous frames, which makes it difficult to locate desired information [7]. Additionally, videos play in one temporal direction, which limits the flexibility and control that users have in how they navigate the content—control and flexibility that they possess when interacting with other GUIs.

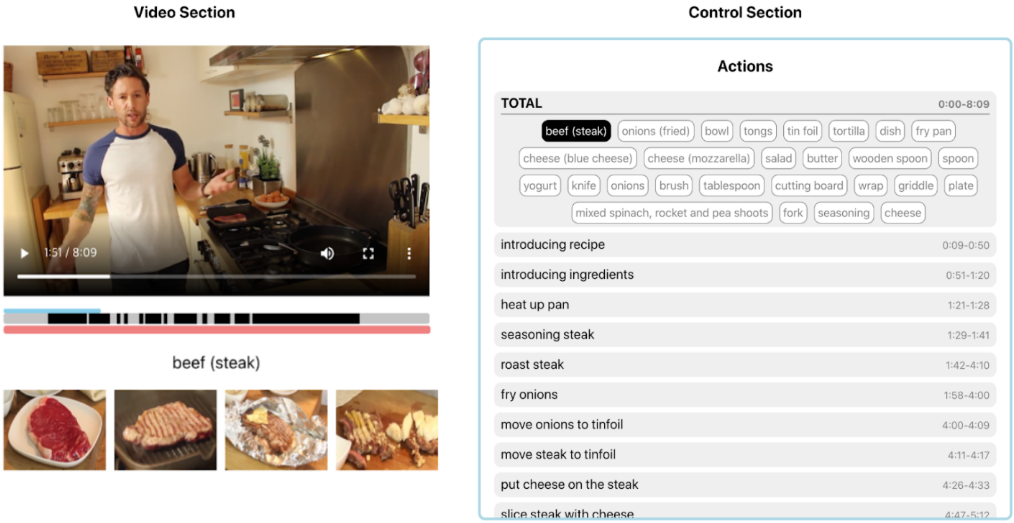
Several approaches have been proposed for video navigation and browsing [15]. Among them, a potential approach to overcoming the aforementioned limitations of videos is to define and identify *informational units* in videos that can serve as points of navigation for users. For example, several systems investigated how transcript keywords [3, 6], frames [9], or concepts [8] contained in the videos could serve as navigation points. Other systems help viewers identify and navigate between meaningful portions of videos by surfacing key information and segmenting videos into chapters [13], scenes [12], intermediate results [7, 11], or spatial locations [16]. Through the investigation of user needs and integration of said needs into video watching systems, this line of work has demonstrated that identifying these informational units can increase the accessibility of the information contained in videos.

While previous work mostly focused on one particular informational unit (e.g., concepts, chapters, intermediate results, spatial locations) to best support their identified user needs, we aim to gain a better understanding of this space: what informational units videos contain, and how each of these help with or are limited when supporting navigation. To do so, we ran a series of studies to investigate the effect of several informational units and understand how these could be expanded on further. We focus on cooking videos as the target domain as they contain complex procedural knowledge involving various objects and actions, and are one of the most widely watched types of videos [5].

Inspired by cognitive psychology theory that posits that humans understand and structure events based on objects and actions [18], we first ran a study to see how users could navigate videos based on the objects and actions extracted from the video. Then, as we learned that the criteria of defining actions should reflect how people perceive the actions, we ran a second study to observe the various ways in which users perceive actions. From this study,

*CHI '22 Extended Abstracts, April 29-May 5, 2022, New Orleans, LA, USA*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn



| | |
|---|---|
| TOTAL | 0:00-8:09 |
| introducing recipe | 0:09-0:50 |
| introducing ingredients | 0:51-1:20 |
| heat up pan | 1:21-1:28 |
| seasoning steak | 1:29-1:41 |
| roast steak | 1:42-4:10 |
| fry onions | 1:58-4:00 |
| move onions to tinfoil | 4:00-4:09 |
| move steak to tinfoil | 4:11-4:17 |
| put cheese on the steak | 4:26-4:33 |
| slice steak with cheese | 4:47-5:12 |

**Figure 1: A video interface used in the first formative study. It lists actions in a video and objects involved in each action.**

we found that participants group actions based on where they happened (i.e., tools or scenes such as "cooking board" or "sink") and describe the dependency information of actions (e.g., "cutting onions" should be done prior to "grinding onions"). Finally, we ran the third study with a video interface that provides the relational information (i.e., spatial and dependency) between actions. We found that it would be more useful to organize video content based on intermediate outcomes and provide consistent time axis between the video and its representation.

With the results from this series of formative studies, we discuss how our findings are related to how humans perceive actions in general. Then, we discuss possible improvements to video interfaces that can be made in the future.

## 2 FORMATIVE STUDY: OVERVIEW

We conducted three consecutive user studies to investigate effective video interfaces by presenting different informational units of videos. In each study, we presented participants with a different version of a video-watching interface and identified challenges and opportunities in using the given information units to navigate and understand the video. We recruited a total of 9 participants (3 female, 6 male, mean age 25), three for each study. Instead of investigating one specific informational unit with all the participants, we chose to do multiple studies to investigate various types of informational units. We picked three cooking videos where a cook is explaining the cooking process.[1] The process involves creating multiple intermediate outcomes as this would increase the complexity of a video and the challenge of identifying information. In our reporting of the findings, we denote participants as P{study_number}-n.

## 3 FORMATIVE STUDY 1: OBJECT & ACTION-BASED VIDEO INTERFACE

Based on the findings of prior work that suggest that objects can be key factors for video navigation [3, 4], we started with an interface that lists the objects that appear in a video to understand how objects are used in video watching. We used Azure Video Analyzer [1], which presents objects extracted from the video through computer vision (CV) and natural language processing (NLP) techniques. However, there were two challenges that made it difficult to observe the user needs by limiting user interaction with the system: 1) unnecessary labels such as minor objects that are unused (e.g., sink), and (2) unorganized labels, as participants found it difficult to find the label they want.

To address these issues, we designed a video interface that filters out unnecessary objects and organizes object labels based on actions (Figure 1). Since the accuracy of object and action labels can affect the user experience significantly, one of the authors manually labeled objects and actions by watching the video, and the other one verified it. We extracted tools and ingredients as objects, and identified a portion of the video as an action if it was done in one place and the cook described them as one unit.

### 3.1 System

The video interface used in this study (Figure 1) was mainly divided into two sections: the information section (left) and the control section (right). The control section lists the actions that appeared in the video, each with the action's name and duration, and labels of shown objects. By clicking on an object label, the time the object appears in the video is marked on the timeline in the information section and manually selected keyframes of the object are provided to show how that object changes over time.

[1] Study 1: https://youtu.be/9ck3CBVbKR4, Study 2: https://youtu.be/ob1mQRnidYQ, Study 3: https://youtu.be/T7icxr899qc

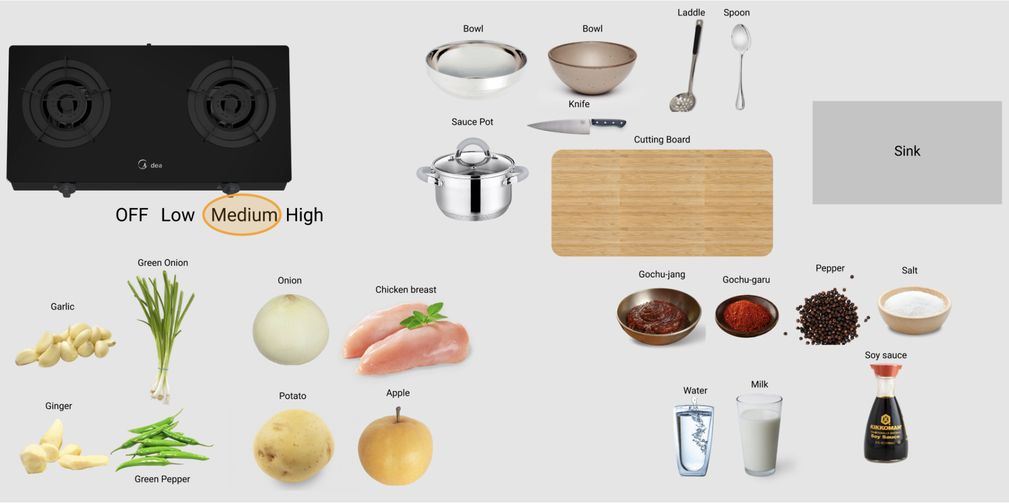


**Figure 2: Virtual cooking space implemented with Figma to simulate following of a cooking video process.**

## 3.2 Study Design

We first gave a brief overview of the system and let the participants explore the system. Then, we asked the participants to skim through the video with the interface for one minute, and asked follow-up questions to verify how well they understood the video. Finally, we asked the participants to imagine following the recipe in the video while using the interface. To make the participants feel like they were actually doing it, we asked users to explain the current step they were doing. Below, we discuss how the object and action information helped participants watch the video and identify challenges.

## 3.3 Findings

*3.3.1 Action and object information.* In the interface, the video was segmented into actions and then segmented again into objects. This hierarchy not only helped participants gain an overview of the video, but it also helped them infer what each action would involve through the constituent objects. Specifically, P1-3 said, *"A list of actions gives a high-level idea of the video before watching it."* P1-2 also said, *"Object labels helped me expect which ingredients, tools, and steps would be included in each action."* It also helped participants revisit a specific part of the video. P1-2 said, *"It was easy to find the seasoning part since I could click on the action name and object in order to get the starting time of the action."*

*3.3.2 Challenges.* Although the system could facilitate navigation compared to only timeline-based interfaces, there was room for improvement regarding the presented actions. P1-3 suggested to carefully determine the granularity of actions and keep the granularity consistent. Also, P1-1 suggested to provide further information on the actions. He said, *"To give both the overview and detail of the video, a list of objects and actions is not enough. Showing relationships or connections between them could improve the interface."*

## 4 FORMATIVE STUDY 2: HOW PEOPLE PERCEIVE ACTIONS ON VIDEOS?

From the first study, we discovered that simply listing actions in chronological order provided limited support, and that how an interface presents actions should follow how users mentally structure actions. To understand how people perceive actions in videos, we ran the second study.

## 4.1 Study Design

In this study, we only provided the participants with a video and asked them to summarize the video in their own way, as this would reveal how users organize, structure, and remember the actions shown in the video. Then, users were asked to simulate the experience of following the video with the summarization they made in a virtual cooking space that we created in Figma (Figure 2). This virtual space contains all the ingredients and tools as movable image objects. We expected that it could better reveal the benefits and limitations of the information structures that the participants created in their summaries. Below, we explain the main findings on how people perceive actions in videos.

## 4.2 Findings

*4.2.1 Grouping actions based on the same surface or tool.* When summarizing the video content, participants grouped individual actions when the actions were done on the same surface (e.g., sink) or with the same tool (e.g., pan, bowl). For example, in a step where the cook cuts several ingredients such as green onions, chili, and pear on a cutting board, P2-1 and P2-2 summarized the step as "ingredients preparation". For the steps of preparing meat and boiling meat, although the actions were both related to the same ingredient, P2-1 and P2-2 distinguished the steps as they were done with different tools.

*4.2.2 Dependency of actions.* When participants were asked to summarize the content, they explicitly represented connections between the steps by indicating which step has to be done prior to

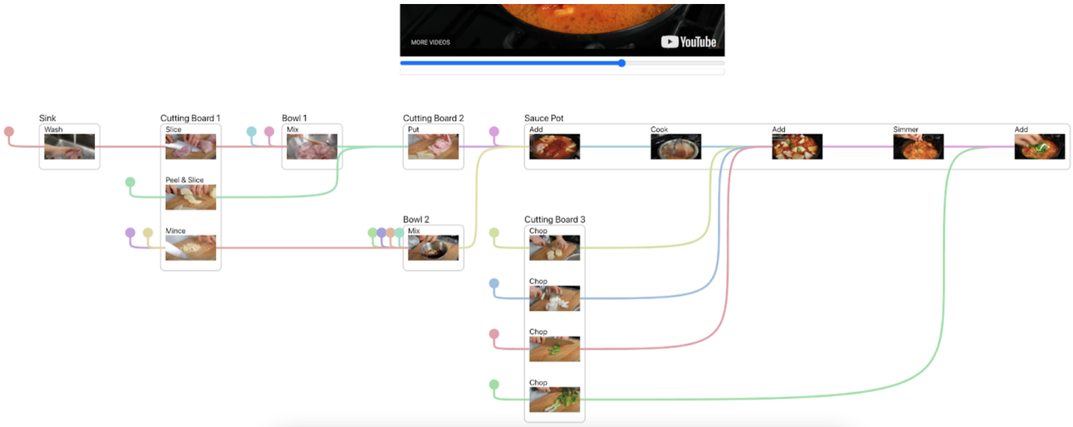


**Figure 3: A video interface used in the third formative study. It groups actions in a video based on where they happen and show dependency information.**

which steps. After writing down the first step, P2-1 illustrated the next two steps that could be done synchronously with two arrows going out from the first step. P2-3 also made labels for each step and made connections between the steps, indicating their order.

## 5 FORMATIVE STUDY 3: RELATION-BASED VIDEO INTERFACE

From the second study, we found out that people perceive the actions in cooking videos based on the surface or tool with which the action happens, and connect the actions by reflecting their dependencies. To investigate how tool-based grouping and visualizing the dependency order would be used in video watching, we implemented a system that reflects relational information of the actions (Figure 3).

### 5.1 System

The interface consists of the main video player and a graph that shows the workflow of the video content (Figure 3). We first segmented the video into several clips based on actions done on an ingredient (e.g., "slicing" a "chicken breast"). Each clip is represented as a thumbnail in the graph. Each ingredient is represented as a small dot and it is connected to a video clip if the ingredient is used in the clip. Then, several clips are grouped in a rectangle if 1) they were done on the same tool (e.g., cutting board) and 2) their actions were not too far away from each other (within one minute). Each clip is connected reflecting their dependencies. Once a user clicks on a video clip, the video player jumps to that part.

### 5.2 Study Design

The task and study setting was the same as the second study, except that participants used the interface above.

### 5.3 Findings

Overall, participants appreciated that they could navigate the video through action units and see the ingredients used in certain steps (P3-2, P3-3). However, there were two main challenges when using the interface, which we discuss below.

*5.3.1 Intermediate outcomes.* The presented interface groups actions performed on the same tool. However, participants wished to see intermediate outcomes and organize the actions based on the intermediate outcomes as well (e.g., sauce, cooked meat). P3-1 said, *"Intermediate outcomes play an important role in understanding the whole process of cooking. It would be better to show how intermediate outcomes, products, or dishes are made."* P3-2 also first sought intermediate outcomes and tried to understand the video based on them.

*5.3.2 Different time axis between the video and the graph.* The presented interface shows steps as a graph, reflecting their dependencies. While P3-3 followed the actions in the dependency order and thought that it would save time, others got confused by the graph due to a mismatched time axis between the video and the graph. P3-2 said, *"It is natural to assume that steps would be organized from left to right chronologically. The overall flow got a little messy for me."* The problem was exacerbated due to the lack of support for locating what the video is showing in the graph. P3-1 said, *"Since the two have different time axes but I could not locate the video in the graph, the matching between the two was difficult."*

## 6 DISCUSSION

From the series of studies, we investigated how objects, actions, and relations between actions provide useful points for video navigation and identified additional challenges that arise from integrating these units. We first discuss how our findings align with human perception theories, and then discuss limitations and possible future directions.

### 6.1 Human Perception Theories

Radvansky and Zacks [14] proposed that, for any event, it is the relational information that provides the unique structure humans use to perceive the event. They categorized relational information into two types: structural relations and linking relations, which are composed of temporal and causal relations. Structural relations specify relations among entities, such as spatial or social relations. Temporal relations specify the chronological order of how events occurred in relation to each other. Lastly, causal relations provide information about the causes and effects of events. Additionally, Catrambone [2] proposed that when people learn a procedure, subgoal information that represents the purpose of a set of steps can guide their learning.

From the first study, we found that temporal relations increased the use of object-based video interfaces. From the second study, we observe that participants specify causal relations between actions when perceiving an event, and also organize them by structural relations that are based on surfaces or tools. From the last study, we observe that participants further want to organize video content based on subgoals.

As videos are a medium that contains a series of events, we could see that the human perception theory of events applies to how users perceive videos. Similar to previous work [16], we could leverage such theories to design more human-centric video interfaces.

## 6.2 Limitations and Future Directions

While our findings align with human perception theories, there are several limitations in our study. First, the number of participants (N=3) for each study might not have been enough to generalize the findings, as there could be various ways of consuming videos. Second, our study focused on cooking videos, which is one of the video domains with the most objects and actions among how-to videos [3]. How people consume videos could be different for videos with fewer objects and actions, such as drawing. Based on the explorations we have done in this work, we plan to iterate on video interfaces with more users and investigate how it can be applied to other domains with a variety of objects and actions.

From our study, we found that it would be more useful to further organize video content based on subgoals [17] in addition to the spatial and dependency relations supported. As such, video content can be grouped in diverse ways. One such way would be a flexible video structure where a video is segmented into multiple snippets and the snippets can be grouped together into multiple levels such as spatial relations and sub-goals. Similar to computational pipelines created to understand and adapt UIs, a multimodal pipeline (CV for video frames and NLP for transcript) could be created to automatically identify useful information units and their relationships to enable this form of video adaptation.

Once we have the structural information of the video, it can be presented in a tree/graph-like diagram or minimap. Users could watch the video by navigating this structure according to their own needs and flexibly jumping between necessary segments, instead of watching the video segments in the order that they were created. By transforming a linear video into a non-linear structure of multiple video snippets, users would be able to interact with and watch the video in a more flexible and fine-grained manner while still retaining a high-level view of the whole video in mind.

## ACKNOWLEDGMENTS

This work 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] Azure. 2022 (accessed February 24, 2022). *Azure Video Analyzer.* https://azure.microsoft.com/en-us/products/video-analyzer

[2] Richard Catrambone. 2012. *Subgoal Learning.* Springer US, Boston, MA, 3230–3233. https://doi.org/10.1007/978-1-4419-1428-6_55

[3] Minsuk Chang, Mina Huh, and Juho Kim. 2021. RubySlippers: Supporting Content-Based Voice Navigation for How-to Videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21).* Association for Computing Machinery, New York, NY, USA, Article 97, 14 pages. https://doi.org/10.1145/3411764.3445131

[4] Pierre Dragicevic, Gonzalo Ramos, Jacobo Bibliowitcz, Derek Nowrouzezahrai, Ravin Balakrishnan, and Karan Singh. 2008. Video Browsing by Direct Manipulation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08).* Association for Computing Machinery, New York, NY, USA, 237–246. https://doi.org/10.1145/1357054.1357096

[5] Google. 2022 (accessed February 24, 2022). *More people are streaming YouTube on their TV screens. Here's what they're watching.* https://www.thinkwithgoogle.com/consumer-insights/consumer-trends/watch-youtube-on-tv/

[6] Juho Kim, Philip J. Guo, Carrie J. Cai, Shang-Wen (Daniel) Li, Krzysztof Z. Gajos, and Robert C. Miller. 2014. Data-Driven Interaction Techniques for Improving Navigation of Educational Videos. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) *(UIST '14).* Association for Computing Machinery, New York, NY, USA, 563–572. https://doi.org/10.1145/2642918.2647389

[7] Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J. Guo, Robert C. Miller, and Krzysztof Z. Gajos. 2014. Crowdsourcing Step-by-Step Information Extraction to Enhance Existing How-to Videos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14).* Association for Computing Machinery, New York, NY, USA, 4017–4026. https://doi.org/10.1145/2556288.2556986

[8] Ching Liu, Juho Kim, and Hao-Chuan Wang. 2018. *ConceptScape: Collaborative Concept Mapping for Video Learning.* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3173961

[9] Justin Matejka, Tovi Grossman, and George Fitzmaurice. 2013. Swifter: Improved Online Video Scrubbing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13).* Association for Computing Machinery, New York, NY, USA, 1159–1168. https://doi.org/10.1145/2470654.2466149

[10] David Mogensen. 2015. I want-to-do moments: From home to beauty. *Think with Google* (2015).

[11] Megha Nawhal, Jacqueline B. Lang, Greg Mori, and Parmit K. Chilana. 2019. VideoWhiz: Non-Linear Interactive Overviews for Recipe Videos. In *Proceedings of the 45th Graphics Interface Conference on Proceedings of Graphics Interface 2019* (Kingston, Canada) *(GI'19).* Canadian Human-Computer Communications Society, Waterloo, CAN, Article 15, 8 pages. https://doi.org/10.20380/GI2019.15

[12] Amy Pavel, Dan B. Goldman, Björn Hartmann, and Maneesh Agrawala. 2015. SceneSkim: Searching and Browsing Movies Using Synchronized Captions, Scripts and Plot Summaries. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (Charlotte, NC, USA) *(UIST '15).* Association for Computing Machinery, New York, NY, USA, 181–190. https://doi.org/10.1145/2807442.2807502

[13] Amy Pavel, Colorado Reed, Björn Hartmann, and Maneesh Agrawala. 2014. Video Digests: A Browsable, Skimmable Format for Informational Lecture Videos. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) *(UIST '14).* Association for Computing Machinery, New York, NY, USA, 573–582. https://doi.org/10.1145/2642918.2647400

[14] Gabriel A. Radvansky and Jeffrey M. Zacks. 2014. *Event Cognition.* Oxford University Press.

[15] Klaus Schoeffmann, Marco A. Hudelist, and Jochen Huber. 2015. Video Interaction Tools: A Survey of Recent Work. *ACM Comput. Surv.* 48, 1, Article 14 (sep 2015), 34 pages. https://doi.org/10.1145/2808796

[16] Anh Truong, Peggy Chi, David Salesin, Irfan Essa, and Maneesh Agrawala. 2021. Automatic Generation of Two-Level Hierarchical Tutorials from Instructional Makeup Videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21).* Association for Computing Machinery, New York, NY, USA, Article 108, 16 pages. https://doi.org/10.1145/3411764.3445721

[17] Sarah Weir, Juho Kim, Krzysztof Z. Gajos, and Robert C. Miller. 2015. Learnersourcing Subgoal Labels for How-to Videos. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15).* Association for Computing Machinery, New York, NY, USA, 405–416. https://doi.org/10.1145/2675133.2675219

[18] Jeffrey Zacks, Barbara Tversky, and Gowri Iyer. 2001. Perceiving, Remembering, and Communicating Structure in Events. *Journal of Experimental Psychology: General* 130, 1 (2001), 29.