VideoWhiz: Non-Linear Interactive Overviews for Recipe Videos

Megha Nawhal* Jacqueline B Lang Greg Mori Parmit K Chilana

School of Computing Science, Simon Fraser University

ABSTRACT

With millions of recipe videos increasingly available online, viewers often face the challenge of browsing through these videos and deciding among different styles of recipe demonstrations and instructions. Although state-of-the-art video summarization techniques using linear presentation formats have been shown to be effective in domains such as surveillance, sports or lecture videos, recipe videos are often more complex and may require a different summarization approach. We first investigated how viewers navigate recipe videos and what information they look for when seeking quick overviews of such videos. Based on our findings, we designed VideoWhiz, a novel interactive video summarization tool that provides a non-linear overview design allowing easy access to the key stages or milestones within the recipe and inter-milestone relationships. VideoWhiz uses a combination of computer vision techniques and an annotation workflow to generate these interactive overviews. Our evaluation showed that viewers found VideoWhiz to be effective and useful in providing quick overviews of recipe videos. We discuss the potential for future work to investigate non-linear overviews for other types of instructional videos and to explore more powerful representations for video summarization.

Keywords: Video navigation; summarization; visualization.

Index Terms: Human-centered computing—Interaction design—Interaction design process and methods—User centered design

1 INTRODUCTION

Free video-sharing websites such as *YouTube*, *Dailymotion*, and *Vimeo* increasingly offer millions of demonstrations for do-it-yourself (DIY) activities, such as cooking, hairstyling, repair, or craft [20]. Recent surveys suggest that cooking videos are among the most watched categories of "How-To" Videos and online food channel subscriptions are on the rise [9,10]. The increasing popularity of cooking videos and the availability of large-scale video datasets has opened up several research opportunities for the fields of Computer Vision (CV) and Human-Computer Interaction (HCI) [1,4].

With the growing repository of online recipe videos shared by both amateurs and professional chefs, viewers often face the challenge of deciding among different styles of recipe demonstrations and instructions [10]. For example, a viewer who searches for "cheesecake recipes" on YouTube will have to inspect several search results consisting of 30 second clips to 30 minute cooking shows for a variety of cheesecakes (e.g., from no-bake cheesecakes to multilayer baked cheesecake recipes). The availability of multiple variations of a recipe for the same dish can make it particularly overwhelming and time-consuming for viewers to select a suitable recipe video to follow [11].

One approach for helping viewers browse different videos more effectively is through the use of automatic video summarization methods. For example, CV-based techniques can be used to create visual overviews for lengthy videos by minimizing redundancy in the video and obtaining a linear breakdown of the content. These techniques have been shown to be useful in particular for screening security footage [16,25] and highlighting events from sports archives [30]. Another class of research augments automatic visual overviews with supporting textual content as anchors for browsing and has been shown to be useful for informational videos (e.g., lecture videos) with minimal visual content [29,35] or step-by-step programming tutorials [33].

Although existing video summarization methods have been shown to be effective in several domains, we know little about how well they suit viewers' needs in navigating recipe videos and in helping them decide which recipe video to follow. In particular, recipe videos are relatively more *complex*: they comprise of multiple steps that have variable duration and have context-dependent semantics. Furthermore, these videos often exhibit substantial intra-class variability in visual appearance, and have steps with multiple dependencies and partial ordering constraints.

In this paper, we investigate how users navigate recipe videos to seek overviews of the content and explore how we can design a user-centered interactive video summarization approach for complex recipe videos. We first carried out a formative study with 13 viewers and analyzed their navigation behaviors with recipe videos. We found that viewers mainly followed a non-linear pattern for seeking recipe overviews and relied on identifying visual *"milestones"* or major components in the recipe demonstrated in the cooking video. Based on these insights, we designed VideoWhiz, a novel interactive tool that presents non-linear overviews of recipe videos by adopting a milestone-driven approach and explicitly shows inter-milestone temporal relationships (Figure 1). To implement *VideoWhiz*, we used automated CV methods in conjunction with a human annotation workflow. Our initial evaluation of VideoWhiz with 10 viewers indicated that VideoWhiz was effective and useful in helping viewers obtain quick overviews of recipe videos.

To summarize, this paper makes the following contributions:

- Empirical results illustrating that viewers adopt a non-linear browsing strategy to navigate recipe videos to obtain an overview of the video's content. In particular, the viewers look for milestones within the recipe and inter-milestone temporal relationships as the recipe video progresses.
- The design and implementation of a novel tool, VideoWhiz, that uses automated techniques combined with a human annotation workflow to create interactive non-linear overviews for recipe videos, making the key milestones easy to access.
- Empirical results that demonstrate that VideoWhiz is effective and useful for obtaining quick overviews of recipe videos.

Overall, our results indicate that the non-linear approach is effective for viewers to seek quick and effective overviews of recipe videos. Although our investigation was limited to the cooking domain, we discuss how our non-linear navigation technique can be extended to other DIY domains that rely on visual attributes of the entities or events involved in the activities. The insights from our work offer several directions for research in HCI to better understand viewers' needs to enhance their interaction with instructional videos and for CV to explore more powerful representations for automated video summarization approaches for DIY domains.

*mnawhal@sfu.ca, jbarlang@gmail.com, mori@sfu.ca, pchilana@sfu.ca

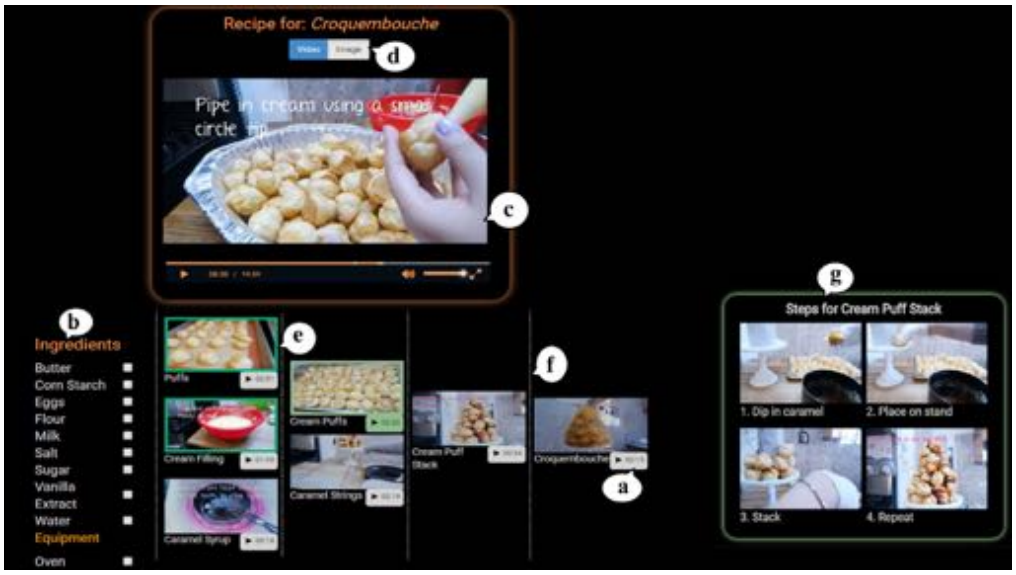


Figure 1: VideoWhiz allows: (a) milestone-driven navigation of the instructional video with access to milestone-specific final outcome and corresponding length of video segment , (b) explicit access to the supplies required for the cooking activity, (c) access to video player interface that can be used to play the milestone-specific segment as well as the full video, (d) toggle option between video mode and the image of the final output of the recipe, (e) explicit access to the milestones (highlighted in dark green in previous level) that are causally related to the currently playing milestone, (f) different levels in the instructions starting from supplies to the final output, and (g) granular steps shown as a storyboard of keyframes on the side with small captions only when the corresponding milestone thumbnail is hovered on.

2 Related Work

Our work builds on prior research focusing on: (1) automated video summarization; (2) video browsing and navigation systems; and (3) visualizations for recipes.

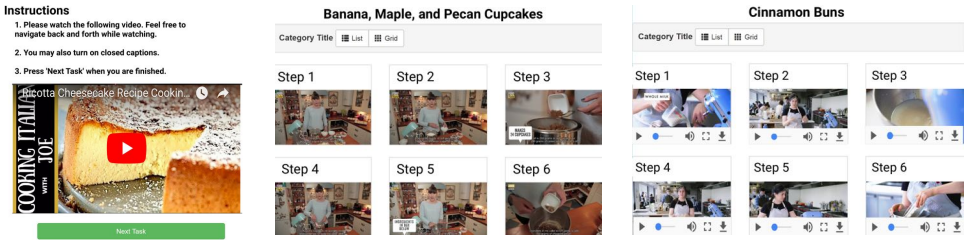
Automated Video Summarization. Automated summarization approaches rely on visual cues to detect salient frames in the video [22] or identify shorter video segments corresponding to key events [16, 23]. Several algorithms leverage subtitles and transcripts in addition to the analysis of the visual information to summarize videos [28]. Some works have created a storyline of events in a video using information from multiple related image sources [34]. Although automated summaries have been shown to be useful for several applications ranging from security footages to sports archives, the algorithms focus on removing redundant content to identify key events and have limited generalizability over videos containing multiple different events such as recipe videos. In contrast, our work illustrates that effective automated understanding of recipe videos calls for representing a video as an aggregate of multiple major components and examining causal relationships between these components of the video.

Video Browsing and Navigation Systems. Conventional video summaries are based on two design principles: (1) obtaining a sequence of important frames in the video known as keyframes [22, 34] or (2) detecting salient video segments known as subshots [8, 15]. Besides these conventional navigational designs, recent approaches have also proposed usage of conversational interfaces for navigation of How-To videos for physical tasks [5]. Systems employing keyframes-based video navigation focus on obtaining major key static instances in videos. These systems facilitate browsing of long entertainment videos [24], supporting textbook-style navigation for lecture videos from Massive Open Online Courses (MOOCs) [26, 35], generating minutes of lengthy egocentric videos [13], and faster detection of events in surveillance videos [15]. Additionally, markers and anchor points based approaches have been used to accelerate the video editing and manipulation process [6, 7, 32]. Toolscape [19] is the most closely related to our tool that provides keyframes-based summary for How-To videos which presents sequence of frames wherein two frames linked by a brief description represent a step in the video.

Video navigation tools using subshots-based summaries aim to generate shorter clips of the video by removing the redundant segments. These summaries accelerate the segmentation of videos to obtain key components in surveillance videos using objects and associated motion trajectories [17, 27] and generating sports video highlights [31]. Moreover, multimodal information such as video transcripts are used to allow video navigation to break down lecture video's content into a linear sequence of key subshots as shown in VideoDigest [29]. Crowdy [33] also uses similar interaction techniques to create linear subgoals for the How-To tutorial videos.

The above summarization systems leverage the linear breakdown of a video's content in the form of either keyframes and subshots to provide a detailed summary of the video's content. However, browsing a detailed summary can be time-consuming for a viewer looking for quick summary of a video before even deciding to watch it end-to-end. In contrast, VideoWhiz adopts a non-linear strategy for obtaining quick overviews focusing on milestones that can be a combination of several events as opposed to the existing linear summaries based on event-level navigation.

Visualizations for Recipes. Although scant, some works have explored ways to improve user interaction with instructional cooking tasks such as investigation of persona-based needs for recipe descriptions [18]. Some works have explored flow-chart based visualizations of a textual descriptions of the recipe [3]. Recent works such as RecipeScape [4] use textual recipe descriptions to help viewers visualize inter-recipe relationships at a large scale. Similar to these works, VideoWhiz is also inspired by the observation that cooking instructions possess a structure that is not necessarily sequential.



(a) **(b)** **(c)**

Figure 2: User interface used during (a) task-specific session of the formative study for participants to browse through the videos; (b) interviews to present keyframes-based overview for recipe R4; (c) interviews to present subshots-based overview for recipe R5

3 Formative User Study

To better understand how viewers navigate recipe videos before deciding which one to watch end-to-end, we carried out an observational user study.

3.1 Method

We selected 5 recipe videos (R1-R5 in Table 1) from the `Cooking_show` category of YouTube-8M dataset [1] to represent a diverse set of videos with varying duration (average duration: 16.47 minutes), and number of steps (24 steps on an average). We recruited 13 participants (6M/7F) with varied levels of routine cooking experience ranging from less than 1 year to 18 years. Our study had two stages, lasting 45 minutes in total: (1): Task-specific observations, and (2): Follow-up interviews.

For the first part of the study, we asked participants to browse three cooking videos (R1-R3 in Table 1) to get an overview of the instructions described in the video. The specific prompt for the participants was to quickly choose one of the three recipes described in the videos which they would prepare for a potluck dinner invitation. The participants were given a time limit of 4 minutes to browse through each of the three videos and decide which one is most appropriate for them.

To minimize learning effects, we developed an interface containing a common video browsing interface embedded for the study provided by the YouTube IFrame Player API [1] along with the instructions as shown in Figure 2(a). Using this interface, we recorded the timestamps of clicks on the video progress bar as participants browsed through the video. The participants had the flexibility to skip around and toggle the captions. Additionally, the participants had access to view thumbnails over the video progress bar. To capture the effect of the appearance of thumbnails on the browsing behavior of the participants, we captured the screen of the participants using Camtasia. We also randomized the order in which the videos were presented to the participants.

During the follow-up interview session, we asked participants to describe: (1) what specific information they were looking for while browsing the recipe videos; (2) why they used (or did not use) captions with the videos; and, (3) what their general strategy is for navigating complex recipe videos before deciding whether or not to prepare a particular dish. We also showed two common designs of video overviews to get initial reactions of the participants on the format of the overview: (1) keyframes for recipe R4 in Table 1 and, (2) subshots for recipe R5 in Table 1. During the video selection process, we consulted 2 experts with more than 20 years of professional cooking experience to provide a step-wise break-down of each of the five recipes. We created the keyframes and subshots based overviews based on the descriptions provided by the experts. We presented the storyboard layout of the keyframes and subshots as shown in Figure 2(b) and (c) respectively and asked participants to explore each interface for 4 minutes (using the same prompt as our main study task described above). We asked participants for their feedback on the strengths and weaknesses of these formats for obtaining a quick overview of recipe videos. During all the study sessions, we encouraged the participants to think-aloud, made audio recordings and used the transcripts in our analysis.

[1] https://developers.google.com/youtube/iframe_api_reference

3.2 Results

Seeking Recipe Milestones and Supplies

When presented with the browsing interface, some of the participants (5/13) started browsing the video by skipping to the end of the video timeline as shown in Figure 3(a) for the visual appearance or "look and feel" of the overall goal or "final output" of the recipe demonstrated in the video. On the other hand, several participants (8/13) started watching the video from the beginning and skipped around to locate the beginning of a different major step in the video. We noticed that these participants often skipped backwards when they noticed a considerable change in the scene content (indicating a potentially new step in the recipe) as shown in Figure 3(b).

During the interviews, all participants agreed that in addition to the explicit visual appearance of the final dish of the recipe, they looked for explicit information about the ingredients and any special equipment requirements. We refer to the ingredients and special equipment required for the cooking activity as *supplies*. Participants' navigation behavior and qualitative feedback indicated that it was important to clearly understand the major stages within the recipe, which we refer to as *milestones*.

Further analysis of the clicks on the video progress bar illustrated that some participants skipped forward (9/13) more quickly as they preferred to skip that part of the video either due to the familiarity with the recipe or the perceived simplicity of that segment. For instance, participant *P10* noted that *"I already know the standard way of making cheesecake crust, I was looking for the ingredients of the custard used in the recipe and if he [the chef] bakes or freezes the cheesecake."* This indicated the need for presenting all the milestones in the instructions before showing corresponding details.

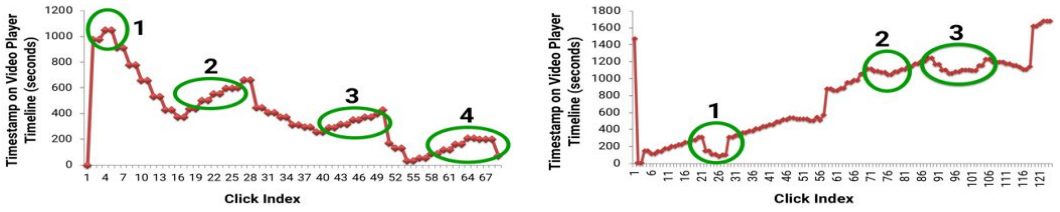
Seeking Inter-Milestone Temporal Relationships

We observed that most of the participants (10/13) tried to associate the step they were watching with the information acquired from the previous steps involved in the cooking activity. For example, participant *P2* explained how she searched for such relationships in the instructions while browsing the video for recipe R2 (Table 1) and described her navigation behavior as *"Where did the green sauce come from?! I just saw the tomato sauce in the video...I was wondering if there is some kind of indicator in the video player showing where did she prepare the green sauce when she is using it in future."* Participants often attempted to identify and establish the temporal relationships between milestones of the recipe as the activity progressed in the video. However, these milestones were not necessarily presented sequentially in the original recipe videos, suggesting non-linear temporal evolution and causal relationships between the milestones of the recipe.

During the interviews, all of the participants further mentioned that being able to explore milestone-specific information such as supplies required to accomplish each milestone would greatly enhance their user experience with the video navigation. Moreover,

Table 1: YouTube videos used for the studies (R1-R5 for formative study and R6-R8 for evaluation). The videos can be obtained from the website using the corresponding Video ID(column 2) in the following URL: `https://www.youtube.com/watch?v=`<Video ID>. The number of steps (#Steps column) was obtained from the breakdown of steps from experts.

Recipe ID	YouTube Video ID	Recipe name	Duration (minutes)	# Steps
R1	X5rxmygZPbo	Italian Almond Biscotti Cookies	20:27	14
R2	RAOglSv09U	Homemade Pizza From Scratch Recipe	17:57	29
R3	u8s7FPuSfEw	Ricotta Cheesecake	29:32	20
R4	LSrN4ZG1JNE	Banana, Maple and Pecan Cupcakes	7:46	28
R5	bqy4XFraBlk	Cinnamon Buns	6:41	28
R6	6ZcXjDMD_y4	Apple Strudel	10:42	22
R7	rAJ2KsH3znc	Homemade Tortellini	11:43	20
R8	A_XkAUqaw9k	Croquembouche	14:34	26
R9	ZQM53dJJKBQ	Apple Crumble Pie	25:35	30



(a) (b)

Figure 3: (a) The navigation patterns of Participant *P11* while browsing recipe video for R3. This indicates the attempts to find the output of each milestone. Labels 1, 2, 3, and 4 denotes instances for cheesecake (final output), baking step, ingredients for custard, and making the crust respectively (b) The navigation patterns of Participant *P2* while browsing the video for recipe R2. The backward skips indicate the viewer's action of tracing the starting point of another milestone. Labels 1, 2, and 3 correspond to finding the starting points of tomato sauce preparation, step for rolling the dough, green sauce preparation for pizza respectively

some of the participants (8/13) mentioned that for complex recipes it is difficult to relate the steps to the final dish and that being able to know the steps that result in a certain part of the final dish would enhance their overall understanding of the recipe.

Seeking Flexibility in Detail: Preference for Combination of Keyframes and Subshots

When presented with the two conventional automated overviews during the interviews, most participants (9/13) preferred the subshots-based overview over keyframes. The key reasons participants did not prefer the keyframes-based overview were that the information was distributed across different images and that the large number of images made it difficult to quickly parse the overall recipe. On the other hand, all of the 9 participants who favored the subshots-based overview wanted to know the outcome of each video segment in the summary before watching it end-to-end.

Furthermore, most of the participants mentioned that all events in the video were not of the same importance to them as they sought to get a quick overview of the recipe. For example, when participants only wanted to get a glimpse of the key events (not necessarily focus their attention), the keyframes-based overview was preferred. Overall, participants showed a preference for a combination of keyframes and subshots to have more control over recipe details.

Seeking enhanced parsing of visual elements: Preference for short captions

We observed that most participants (9/13) turned video captions off as the long captions were perceived to be "distracting"; but, the captions were helpful for participants in understanding the key supply requirements needed for the cooking activity. Moreover, almost all participants (12/13) mentioned that having a short textual description for both video segments and images would reduce the time consumed in inferring the content of individual image or segment. This would allow the viewer to focus on the milestones that require more attention. We refer to any video segment or image in the recipe overview as a *visual element* hereafter.

3.3 Design Implications

Based on our findings, we synthesized 5 key design implications for creating quick overviews for recipe videos.

- **D1: Disintegrating recipe into milestones.** Viewers should be able to access visual information about the key milestones or major intermediate stages in lengthy recipe videos. Moreover, viewers should be able to easily access the overall goal and supplies required for the activity.
- **D2: Connecting the recipe milestones.** Viewers should be able to quickly visualize the temporal evolution of the instructions being demonstrated in the video. In addition, viewers should be able to visualize causal relationships between milestones such as the ones must be completed before executing a particular milestone and the ones that can be executed independently.
- **D3: Describing milestones in hybrid overviews.** Viewers should be able to see hybrid overviews consisting of keyframes and subshots to provide an overview of the milestones and detailed description consisting of a single or multiple segments from the video.
- **D4: Captioning milestones.** Viewers should have access to concise captions associated with each visual element to aid them in allocate their navigation time based on their needs.
- **D5: Exploring milestone-level information easily.** Viewers should be able to query the milestones based on the supplies required in the recipe (bottom-up) or the relation of milestones to the final appearance of the final dish (top-down) to explore milestone-specific content as well as the overall content efficiently.

4 VIDEOWHIZ: SYSTEM DESCRIPTION

Based on our key design implications, we developed VideoWhiz to provide non-linear interactive overviews for recipe videos.

4.1 System Design

Milestone-based Non-linear Timeline

To facilitate milestone-based non-linear navigation (design implication D1 & D2), VideoWhiz presents a non-linear timeline-based overview interface (Figure 1). This timeline appears under the video player where each milestone is described using a short tag (design

implication D4) and the thumbnails show the final outcome of the corresponding milestone (design implication D1) as shown in Figure 1a along with the duration of the video segment of respective milestone. When a viewer chooses to play the video segment using the play button next to the corresponding thumbnail in the timeline (Figure 1e), the system displays the start and end point of the video segment for the milestone on the video player's progress bar. VideoWhiz shows only the start and end points and automatically skips to the next subshot if the video segment has multiple subshots.

The timeline is divided into levels starting from list of supplies in the first level to the milestone corresponding to the final output present in the last level as shown in Figure 1f. For example, Figure 1 shows the thumbnails for the seven milestones in the recipe for *"croquembouche"* along with the supplies required for the recipe on the left and image of *croquembouche* in the last level. Other milestones are arranged in the intermediate levels between the supplies and final output based on the evolution of the set of instructions demonstrated in the video. The key features of the timeline addressing design implication D2 are: (1) the parallel-placed milestone thumbnails, and (2) explicit indication of milestone *precursors*, i.e., the milestones that need to be executed before the given milestone. The milestones that are in the same level and are parallel-placed can be executed independently. The execution of milestones in a level would involve all or some of the milestones in the previous levels indicated by highlighting the milestone's precursors in the previous levels on hover over the thumbnail of viewer's interest in the timeline, see Figure 1e. Consider the overview for *"croquembouche"* recipe in Figure 1, milestones in the first level, namely, *puffs, cream filling*, and *caramel syrup* can be prepared independently using some or all of the ingredients in the list on the left of the timeline. Similarly, *cream puffs* in the second level requires *cream filling* and *puffs* to be prepared. When the viewer chooses to play a video segment corresponding to the milestone, the thumbnail corresponding to the milestone currently playing and precursor milestone remain highlighted while rest of the thumbnails are greyed out.

Structured Overview with Hybrid Format
To support design implication D3, VideoWhiz introduces a hybrid overview design that consists of keyframes and subshots for each identified milestone. When a viewer hovers over a particular milestone thumbnail in the storyboard, an additional keyframes-style overview pop-up appears on the side of the screen (as shown in Figure 1f). Each of the keyframes in the pop-up box also have text captions (design implication D4). Moreover, a viewer can also play the video segment for a more detailed description of the step by clicking on the play button next to the thumbnail.

Querying the Timeline using Visual Components
As per our design implication D5, users should be able to look for milestone-specific information from our system. To provide a quick top-down visual understanding of the recipe to the viewer, we breakdown the image of the dish into different components which illustrate their correspondence to the milestones.

VideoWhiz explicitly shows the image of the overall output of the recipe demonstrated in the video as the default view above the milestone-based timeline. The image of the overall output is divided into the visually separable components using an imagemap. Each area in the imagemap is highlighted with a border when the viewers hovers on it (Figure 4) and the milestones in the timeline that are involved in the selected (visual) component of the overall output are highlighted with other milestones being greyed out. For example, in the overview for the recipe of *"apple crumble pie"* shown in Figure 4, when the image region corresponding to *crumble* was selected the milestones *crumble mixture* and *final assembly* are highlighted showing that these milestones were involved in making the *crumble* component of the pie. Additionally, the viewer always has access to the video player using the toggle button (Figure 1d).

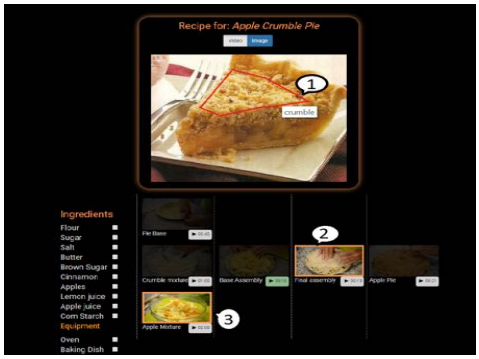


Figure 4: Querying the visual component *crumble* denoted as (1) in the recipe R9 and the relevant highlighted milestones (2) and (3)

Filtering Milestones by Supplies
As per our design implication D5, to provide a quick bottom-up understanding of the recipe, Videowhiz provides explicit access to the information about the supplies required for the cooking task, (*i.e.*, cooking ingredients and special equipment). In our system, the supplies required are displayed next to the timeline for viewers to have easy access to the information while browsing the video (Figure 1b). The check-box next to each item in the supplies section of the interface can be clicked to query the timeline to identify milestone(s) that involve usage of the selected supply item.

4.2 System Implementation

To implement VideoWhiz, we used a machine-assisted human annotation workflow approach: (1) CV and NLP based computational pipeline; and, (2) human annotation pipeline to refine the automatically-generated annotations.

Computational Pipeline
To detect candidate milestones in a recipe video, we used an image similarity-based approach. We extracted the individual frames at 5 fps and maintained the exact timestamp of the frame in the video as the associated metadata. The frames were mapped to a vector representation using the penultimate layer of the neural network ResNet-50 [12] pretrained on ImageNet dataset [2]. Subsequently, we used Euclidean distance between the vector representations of the frames as the similarity measure between the consecutive frames and appended to a similarity vector. We detected the frames corresponding to the points of local maxima in the similarity vector. These frames corresponding to the local maxima were the candidate scene-transition frames in the video. Additionally, we obtained the audio transcriptions and parsed the captions corresponding to the candidate scene-transition points using the NLTK POS tagger [3] to obtain the noun and verb phrases as the candidate supplies and actions respectively.

However, the computational pipeline did not incorporate the semantic understanding of the scenes and captions leading to inaccuracies. For instance, some candidate scene-transition frames may not be relevant to the activity demonstrated in the video, such as the chef talking about a particular brand of the olive oil she uses while making pasta. Therefore, we used a human annotation pipeline to refine the automatically-generated annotations.

Human Annotation Pipeline
The annotation pipeline that we devised consisted of two stages: the

[2]https://download.pytorch.org/models/resnet50-19c8e357.pth
[3]http://www.nltk.org/book/ch05.html

goal of the first stage was to determine the recipe milestones (using our definition above), and the second stage obtained the details of each of these milestones. To make the implementation scalable, we modularized the annotation pipeline such that the second stage tasks (designed for collecting details) could be conducted independently after obtaining the first stage annotations (described below).

(a) (b)

Figure 5: Interface for Stage 1-Task 1

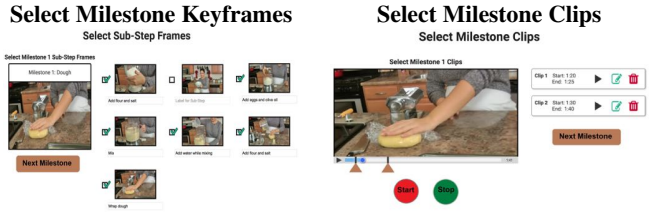
• *Stage 1-Task1: Identifying milestones.* The annotators were asked to (a) click on the checkbox next to the relevant images representing milestones in the set of candidate scene-transition frames Figure 5(a) and, (b) provide a brief caption (less than 5 words) in the textbox below the selected images Figure 5(b).

• *Stage 1-Task 2: Identifying inter-milestone connections*. The annotators clicked on the checkbox next to the images of the milestones (right) that are connections of the given milestone (left) as shown in Figure 6.

Figure 6: Interface for Stage 1-Task 2

• *Stage 2-Task 1: Summarizing key steps for milestones.* The annotators clicked on the checkbox next to the images relevant to the milestone (left) and provided a description of less than 5 words in the textbox below each relevant image (see Figure 7(a)).

• *Stage 2-Task 2: Refining video segments for milestones.* The annotators could play the candidate video segment using the start button. Our tool collected the timestamps of these segments that were considered relevant until the stop button was pressed; the segments could be edited and replayed to verify the annotations (right) (see Figure 7(b)).



(a) (b)

Figure 7: Interface used for Stage 2-(a)Task 1, (b)Task 2

• *Stage 2-Task 3: Identifying supplies & associated milestones.* The annotators could click on the checkbox next to the images (Figure 6) corresponding to the milestones where the given ingredient (shown on the left replacing the left image in Figure 6).

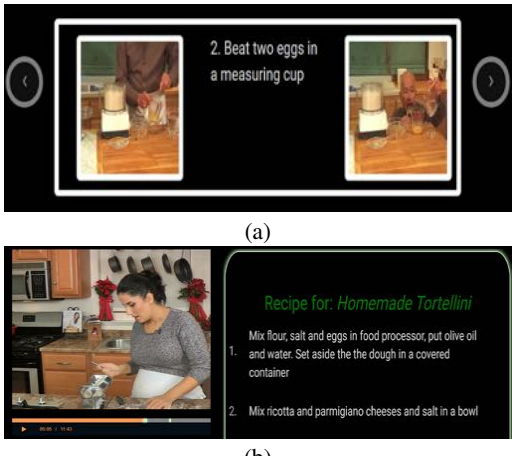
• *Stage 2-Task 4: Identifying output's visual components & associated milestones.* The annotators used a web-based ImageMap generator [4] to obtain the outlines for visually separable components in the image of the final output of the recipe. They marked the main dish for the recipes where the image of final output did not contain visually separable components. Subsequently, annotators could click on the checkbox next to the images (Figure 6) corresponding to the milestones involved in execution of the given component of the dish (shown on the left replacing the left image in Figure 6).

[4] www.image-map.net

Execution and Initial Validation: Before performing each of the annotation tasks, the annotators were asked to browse through the corresponding recipe video and were informed that they would have access to the video during the task. We recruited 10 annotators in total: 8 annotators with no or minimal cooking experience referred to as *regular annotators*, and 2 annotators with more than 10 years of cooking experience referred to as *expert annotators*. For both stages, each task involved 2 regular annotators. Subsequently, 1 expert annotator was assigned to each task to validate and merge the annotations obtained from regular annotators in Stage 1 and Stage 2. We carried out the annotation instances for 4 cooking videos (R6-R9 in Table 1). For all the videos, we observed an overlap of 77% and 80% (averaged over 4 videos) in the annotations for the two regular annotators of Task 1 and Task 2 in the Stage 1 respectively. We also noted significant overlap of 72-80% (averaged over 4 videos) for the 4 tasks in the Stage 2 annotations from the regular annotators.

5 Evaluation of VideoWhiz

Our main goal in designing VideoWhiz was to enhance the video navigation experience of viewers who intend to get a quick overview of a recipe video before deciding to watch it end-to-end. To measure the effectiveness and usefulness of the non-linear overviews created using VideoWhiz, we carried out an observational user study. As part of our evaluation, we also compared VideoWhiz with two state-of-the-art enhanced video overview designs that have been used in other applications. These overview designs included: (1) text-augmented keyframes-based linear summarization systems such as Toolscape [19], and Swifter [24], (2) text-augmented subgoal-driven subshots-based linear summarization systems such as Crowdy [33] and VideoDigest [29]. For simplicity, we refer to these systems as enhanced keyframes-based overviews and enhanced subshots-based overviews respectively hereafter. For both of these two interfaces, we developed our version as shown in Figure 8 with annotations obtained from the experts based on the definitions and instructions provided in the respective papers.



(a)

(b)

Figure 8: Our implementation of the user interfaces for (a) enhanced keyframes-based overview system [19] and (b) enhanced subshots-based overview system [29,33] used for comparison with VideoWhiz

5.1 Study Setup and Protocol

For the evaluation, we used three cooking videos for recipes R6, R7 and R9 listed in Table 1 (average duration: 16 minutes). We cre-

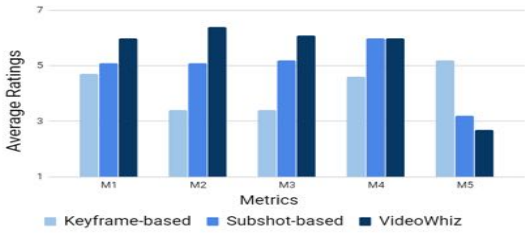


Figure 9: Average ratings (on a scale of 7) for the three different overview systems with respect to the metrics indicating the effectiveness and usefulness of VideoWhiz (higher the better for M1-M4, lower the better for metric M5)

ated enhanced keyframes-based overview, enhanced subshots-based overview and non-linear interactive overviews using Videowhiz for R6, R7 and R9 respectively. To provide a brief tutorial to the participants, we used the overviews created for recipe R8 (Table 1) for each of the three overview interfaces.

We conducted the study with 10 participants (4M/6F) with varied levels of routine cooking experiences ranging from less than an year to 21 years of cooking experience. All the participants for the evaluation study were different from the formative study (Section 3). All of the participants tried out each of the three overview systems for approximately 2 minutes where the prompt was same as the formative study, (*i.e.*, to quickly decide on preparing one of the three dishes for a potluck dinner based on the recipe description in the overviews). To balance out learning effects, we randomized the order in which the three overviews were presented to the participants. We encouraged the participants to think-aloud during the sessions.

After browsing through each of the overview interfaces, the participants filled out a post-task questionnaire with relative ratings (on 7-point Likert scale) comparing the usefulness and effectiveness of each overview system. We defined effectiveness as the extent to which the overview serves the recipe overview needs of the participant with respect to the following metrics.

- Ease of finding required information about the recipe using the system (M1)
- Ease of locating major components of the recipe (M2)
- Preference to reuse the system for an overview of another recipe video (M3)
- Preference to use the system to follow the cooking task step-by-step as demonstrated in the video (M4)
- Need to browse the video even after using the overview (M5)

Each session took approximately 30 minutes, including the tasks and a brief follow-up interview of 8-10 minutes about participants' experiences with each of the three overview formats.

5.2 Results

Overall, our findings indicate that participants found VideoWhiz to be effective in providing quick overviews of recipe videos (participant ratings are summarized in Figure 9). All of the 10 participants either agreed or strongly agreed that VideoWhiz allowed them to quickly find the information they needed from the recipe (M1) and that the proposed overview interface presented all the major components of the recipe (M2). Most participants (8/10) preferred to reuse the VideoWhiz interface for obtaining overviews for other similar videos (M3) and only 1/10 was inclined to watch the full video after browsing the interface (M5). On the other hand, the response for the enhanced subshots-based and enhanced keyframes-based overview interfaces were mixed. All of our 10 participants either somewhat agreed (7/10) or agreed (3/10) for metrics M1, M2, and M3 when viewing the enhanced subshots-based overview. Moreover, most participants (6/10) noted they would not watch the full video after browsing through the interface. The enhanced keyframes-based overview also was ranked low, with several (7/10) participants disagreeing about the effectiveness of enhanced keyframes overviews (M1, M2, M3 and M5). Also, the enhanced keyframes-based overview interface was less popular for reuse as most participants (8/10) indicated that they would not want to use the overview format again.

In addition to evaluating the effectiveness of the system for obtaining quick overviews, we also assessed the perceived effectiveness of the different overviews for following a task step-by-step (M4). We observed that all of our 10 participants agreed or strongly agreed that they could use enhanced subshots-based overview and VideoWhiz to perform a task step-by-step. In particular, participants appreciated the flexible control over viewing recipe details in VideoWhiz. In contrast, the enhanced keyframes-based overview was perceived to be less effective for following steps.

Furthermore, the qualitative feedback from participants in the follow-up session corroborated the questionnaire results. Most participants found VideoWhiz overviews to be useful not just for getting a quick overview of the recipe, but also for inspecting specific milestones or steps involved in the instructions in more detail. The participants indicated that the structured format could especially be useful for recipes that involve multiple high precision steps (*e.g* rolling, baking) and have multiple steps related to each other. In contrast, most of the participants (9/10) were less enthusiastic about the enhanced keyframes-based design where the information about recipe steps could be distributed over several visual elements and difficult to locate within the keyframes due to large number of images present. Also, 8/10 participants mentioned that subshots-based overview provides a shorter description of the steps but they would still need to watch the video which made the browsing arduous unlike VideoWhiz where the information is explicitly presented.

Despite the overall positive feedback, some participants noted that additional information within the overview (e.g., baking times, cooking temperatures, exact quantities of ingredients) would make VideoWhiz even more useful if it were to be used as an instructional aid (in addition to being a summarization tool).

6 Discussion and Conclusion

In this work, we have investigated how viewers navigate complex recipe videos as they seek quick overviews and we have introduced VideoWhiz, a novel summarization approach that provides interactive milestone-based non-linear overviews. We now discuss some of the limitations of our work and opportunities for future work in HCI and CV.

Generalizability of Milestone-Driven Design. Although we focused our investigation on the cooking domain, the underlying design principles of VideoWhiz can be generalized to other DIY domains where the description of the visual attributes of events or entities is useful for viewers in getting quick overview of the underlying DIY activity (*e.g.* craft, repair, home management, or hairstyling). For example, the task of assembling a chest of drawers from IKEA would involve milestones such as assembly of the drawer frame and separate drawers that form precursors to the final assembly. Furthermore, our initial evaluation suggested that participants found VideoWhiz to be as effective as the text-augmented subshots-based system for following the demonstrated activity step-by-step. Further examination of the overviews for execution of a DIY activity using the video can improve the applicability of the proposed system to creating effective learning tutorials.

Opportunities for Automated Video Understanding. In our system design, the low accuracy of image similarity based algorithms used for detection of scene transitions led to the need for human intervention in obtaining recipe overview needs for VideoWhiz. While recent approaches [14] can improve the accuracy of the candidate milestones detection, the crowdsourcing workflow can be time-consuming and inefficient for complex DIY videos. With the evolution of CV algorithms addressing semantic

understanding of human activity videos, there is potential to increase the machine assistance in the annotation workflow and reduce the annotation cost, as shown by Branson et al [2].

Moreover, the empirical results from our formative study indicated that viewers' navigation patterns reveal their intention to seek milestones within DIY tasks. Further examination of the viewers' browsing patterns could provide algorithmic insights for user-centered saliency models for the timeline of complex DIY videos using the sparse annotations for events of interest.

Scalability of the Implementation Requirements. To obtain implementation requirements for our system, we relied on getting a few annotators to perform several annotation tasks and observed considerable overlap in their annotations. While we used this pipeline for prototype design, the annotation approach can also be adjusted depending on the system deployment settings. For instance, the implementation requirements could be curated by the video author when used in a video authoring system. In other cases, video hosting websites can use our modularized annotation approach for obtaining annotations from the viewers or third party annotators in a scalable manner. Additionally, to improve the inter-rater reliability among annotators, it would be useful to explore the design of collaborative annotation interfaces [21].

Visualization of DIY Videos at Scale. We observed that viewers found structured non-linear design provided by VideoWhiz to be effective in providing quick overviews of recipe videos. This design could potentially be used to compare various videos pertaining to a single activity at a large-scale as shown earlier in systems such as RecipeScape [4]. Moreover, such large scale visualizations could promote creativity in DIY domains by allowing the viewers to draw inspirations from different ways of performing a DIY activity and create innovative variants of their own.

Customization of Detail in Overviews. During the annotation process, we noted that annotators sometimes differed in the partitioning of some milestones. For instance, the preparation of batter for the recipe of Cupcakes (R5) involved several dry ingredients such as flour and salt followed by mixing of wet ingredients such as eggs, and 1 of the 4 annotators marked the mixing of dry ingredients and wet ingredients as two separate milestones. Extending from these observations, it would be valuable to further characterize viewers based on their needs and level of expertise to make the summaries more effective and useful, which we hope to address in future.

REFERENCES

[1] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[2] S. Branson, G. Van Horn, and P. Perona. Lean crowdsourcing: Combining humans and machines in an online system. In *Proc CVPR'17*.

[3] L. Buykx and H. Petrie. Recipe sub-goals and graphs: an evaluation by cooks. In *Proceedings of the ACM multimedia 2012 workshop on Multimedia for cooking and eating activities*.

[4] M. Chang, V. M. Hare, J. Kim, and M. Agrawala. Recipescape: Mining and analyzing diverse processes in cooking recipes. In *Proc CHI'17*.

[5] M. Chang, A. Truong, O. Wang, M. Agrawala, and J. Kim. How to design voice based navigation for how-to videos. In *Proc CHI'19*.

[6] P.-Y. Chi, S. Ahn, A. Ren, M. Dontcheva, W. Li, and B. Hartmann. Mixt: automatic generation of step-by-step mixed media tutorials. In *Proc UIST'12*.

[7] P.-Y. Chi, J. Liu, J. Linder, M. Dontcheva, W. Li, and B. Hartmann. Democut: generating concise instructional videos for physical demonstrations. In *Proc UIST'13*.

[8] M. Ellouze, N. Boujemaa, and A. M. Alimi. Im (s) 2: Interactive movie summarization system. *Journal of Visual Communication and Image Representation*, 21(4).

[9] Google. I want-to-do moments: From home to beauty. https://www.thinkwithgoogle.com/marketing-resources/micro-moments/i-want-to-do-micro-moments/, 2015.

[10] Google. Millenials eat up youtube food videos. https://www.thinkwithgoogle.com/consumer-insights/millennials-eat-up-youtube-food-videos/, 2015.

[11] M. J. Halvey and M. T. Keane. Analysis of online video search and sharing. In *Proceedings of the eighteenth conference on Hypertext and hypermedia*.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc CVPR'16*.

[13] K. Higuchi, R. Yonetani, and Y. Sato. Egoscanning: Quickly scanning first-person videos with egocentric elastic timelines. In *Proc CHI'17*.

[14] D.-A. Huang, J. J. Lim, L. Fei-Fei, and J. C. Niebles. Unsupervised visual-linguistic reference resolution in instructional videos. *CVPR'17*.

[15] D. Jackson, J. Nicholson, G. Stoeckigt, R. Wrobel, A. Thieme, and P. Olivier. Panopticon: a parallel video overview system. In *Proc UIST'13*.

[16] Z. Ji, Y. Su, R. Qian, and J. Ma. Surveillance video summarization based on moving object detection and trajectory extraction. In *Proc ICSPS'10*, vol. 2.

[17] T. Karrer, M. Weiss, E. Lee, and J. Borchers. Dragon: a direct manipulation interface for frame-accurate in-scene video navigation. In *Proc CHI'08*.

[18] S. J. Kerr, O. Tan, and J. C. Chua. Cooking personas: Goal-directed design requirements in the kitchen. *International Journal of Human-Computer Studies*, 72(2).

[19] J. Kim, P. T. Nguyen, S. Weir, P. J. Guo, R. C. Miller, and K. Z. Gajos. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proc CHI'14*.

[20] B. Lafreniere, T. Grossman, and G. Fitzmaurice. Community enhanced tutorials: improving tutorials with multiple demonstrations. In *Proc CHI'13*.

[21] W. S. Lasecki and J. P. Bigham. Interactive crowds: Real-time crowdsourcing and crowd agents. In *Handbook of human computation*.

[22] Y. J. Lee, J. Ghosh, and K. Grauman. Discovering important people and objects for egocentric video summarization. In *Proc CVPR'12*.

[23] Y. J. Lee, J. Kim, and K. Grauman. Key-segments for video object segmentation. In *Proc ICCV'11*.

[24] J. Matejka, T. Grossman, and G. Fitzmaurice. Swifter: improved online video scrubbing. In *Proc CHI'13*.

[25] A. H. Meghdadi and P. Irani. Interactive exploration of surveillance video through action shot summarization and trajectory visualization. *Proc TVCG'13*, 19(12).

[26] T.-J. K. P. Monserrat, S. Zhao, K. McGee, and A. V. Pandey. Notevideo: facilitating navigation of blackboard-style lecture videos. In *Proc CHI'13*.

[27] C. Nguyen, Y. Niu, and F. Liu. Video summagator: an interface for video summarization and navigation. In *Proc CHI'12*.

[28] R. Panda, A. Das, Z. Wu, J. Ernst, and A. K. Roy-Chowdhury. Weakly supervised summarization of web videos. In *Proc ICCV'17*.

[29] A. Pavel, C. Reed, B. Hartmann, and M. Agrawala. Video digests: a browsable, skimmable format for informational lecture videos. In *Proc UIST'14*.

[30] Y. Takahashi, N. Nitta, and N. Babaguchi. Video summarization for large sports video archives. In *Proc ICME 2005*.

[31] A. Tang and S. Boring. Epicplay: Crowd-sourcing sports video highlights. In *Proc CHI'12*.

[32] A. Truong, F. Berthouzoz, W. Li, and M. Agrawala. Quickcut: An interactive tool for editing narrated video. In *Proc UIST'16*.

[33] S. Weir, J. Kim, K. Z. Gajos, and R. C. Miller. Learnersourcing subgoal labels for how-to videos. In *Proc CSCW'15*.

[34] B. Xiong, G. Kim, and L. Sigal. Storyline representation of egocentric videos with an applications to story-based search. In *ICCV'15*.

[35] K. Yadav, A. Gandhi, A. Biswas, K. Shrivastava, S. Srivastava, and O. Deshmukh. Vizig: Anchor points based non-linear navigation and summarization in educational videos. In *Proc UIST'16*.