Tools to support expository video capture and access

Scott Carter · Matthew Cooper · John Adcock · Stacy Branham

Abstract Video content creators invest enormous effort creating work that is in turn typically viewed passively. However, learning tasks using video requires users not only to consume the content but also to engage, interact with, and repurpose it. Furthermore, to promote learning with video in domains where content creators are not necessarily videographers, it is important that capture tools facilitate creation of interactive content. In this paper, we describe some early experiments toward this goal. Specifically, we describe a needfinding study involving interviews with amateur video creators as well as our experience with an early prototype to support expository capture and access. Our findings led to a system redesign that can incorporate a broad set of video-creation and interaction styles.

Keywords How-to · Tutorial · Lecture · Video · Mobile · Capture and access

1 Introduction

The ways in which we learn and share knowledge with others are deeply entwined with the technologies that enable the capture and sharing of information. As face-to-face communication becomes supplemented with rich media – textual books, illustrations and photographs, audio, film and video, and more – the possibilities for knowledge transfer expand. In particular, one of the latest trends to emerge amid the growth of Internet sharing and pervasive mobile devices is the mass creation of online expository videos, including how-to, tutorial, and lecture videos. In this work, we explore how we can augment

Scott Carter, Matthew Cooper, John Adcock
FX Palo Alto Laboratory, Inc.
E-mail: carter,cooper,adcock@fxpal.com

Stacy Branham
Center for HCI, Virginia Tech
E-mail: sbranham@vt.edu

the video capture and access processes for both lightweight as well as more procedural expository content.

While past work has shown that video is not always the best presentation format for all learning tasks [18], graphics that *show* how to accomplish a task improve understanding beyond textual descriptions [10]. For some tasks video has been shown to be particularly helpful beyond static graphics [8, 21]. This is intuitive since some tasks involve a gradual transition that is difficult to show in static photos (for example, fluffing egg whites). Other tasks might require some other form of multimedia feedback (for example, playing a tin whistle). Video can also help coordinate a series of steps into a global action described statically. For example, the act of kicking a football can be shown as a series of static shots: lining up the foot, striking the ball at a particular spot, following through, etc. But without seeing these individual elements combined in one swift strike it can be difficult to know what the composite end result should realistically look like. Furthermore, video does not preclude integrating static content – many video editing tools support the integration of static photos that can be "played" for some period of time within the video. For these reasons, we focus on video-based support for expository content.

However, video alone cannot support all of the tasks involved in creating tools to support learning. Often, expository video involves a more definite progression of steps or important points than other genres. We hypothesize it may be useful to expose this structure at the interface level. Some tasks may also require supporting documentation such as text, high resolution photos, schematics, audio clips, etc.

To test our intuition that video augmented with bookmarks and multimedia annotations can enhance the capturing and accessing processes for expository video, we took a two-pronged approach: one focusing exclusively on understanding current practice and one experimenting with an early prototype. Our goal was to understand both latent issues with off-the-shelf tools and also to gain a sense of how likely users would be to adopt novel expository capture and access tools. Our findings suggest that tools should support a broad set of creation styles with a unified access and annotation interface, the design and construction of which we describe at the conclusion of the paper.

2 Tool requirements and design

Our approach to uncovering the requirements necessary for tools to support expository video content involved 1) understanding past work investigating the role of media in learning and knowledge transfer; 2) conducting first-hand, participatory observations of the use of off-the-shelf capture and access tools; and 3) developing and piloting a mobile application.

2.1 Lessons from the literature

Eiriksdottir and Catrambone conducted an extensive review of instructions for procedural tasks that has particular applicability to expository video content [7]. The authors suggest that specific procedural instructions grounded with realistic examples and sparse use of the more general principles involved in the task all contribute to better initial task performance but poor learning and transfer to other tasks. On the other hand, a greater emphasis on principles combined with "fading", or relating specific examples and instructions to higher-level concepts, can help transfer and learning.

In many cases, users of how-to or tutorial videos will need to fix an object without particularly needing or wanting to learn about general principles – for example, when fixing their printer. Thus it is critical that tools support initial performance, implying a focus on step-by-step instructions. Other work has shown that higher quality examples correspond with better task performance [20] and that learning can improve when complemented with video-based examples in particular [15]. Coupled with Clark's and Mayer's finding that multimedia is especially useful for "learners who have low knowledge of a domain," [6] this work suggests that tools for creating tutorial and how-to video should support links to concrete examples and complementary multimedia materials. However, it is equally important that the tool make it possible for users to develop knowledge transferable to other tasks and domains. To support this more general knowledge, tools should facilitate users actively navigating [22] as well as annotating and editing to develop their own interpretations of video content [29,3]. Zhang et al. found that interactive video in particular "achieved significantly better learning performance" than non-linear video because 1) content can be repeated; 2) the interface enables random access, which "is expected to increase learner engagement" and allows the user to control the pace of learning; and 3) it can increase learner attentiveness [30].

Overall, then, past work suggests that interactive video complemented with rich multimedia materials and specific examples can help users both complete short-term tasks as well as potentially develop transferable knowledge. This principle guided our pilot designs.

2.2 Understanding needs first-hand

Design problems uncovered first-hand often vary considerably from those that users report in surveys [19]. One approach to address this is needfinding, a core user-centered design method in which researchers observe and optionally engage participants while they complete a task of interest with the goal of identifying opportunities for improvement [19]. Needfinding will often go beyond passive observation to include post-hoc interviews, contextual inquiry, or other related techniques [9,4,11]. By observing users' actual practices, and asking them questions in situ, designers are able to gain a first-hand understanding of a task's key pragmatic issues.

Pseudonym	Age	How-to genre	Number of how-tos published
Anne (P1)	early 60s	cooking, cleaning and organization, handcrafted earrings	4
Bruce (P2)	mid 20s / late 20s	piano tutorials	34
Carman & Craig (P3 and P3.5)	mid 20s	cooking	5
Derek (P4)	late 30s	software, electronics, musical instruments	8
Elena (P5)	early 30s	gardening, handcrafted earrings, cooking	4
Faith (P6)	mid-30s	beekeeping, gardening, cooking	39
Gary & Greg (P7 and P7.5)	60s / 40s	software tutorials	8

Table 1 Study participants.

In our needfinding work we focused on a specific class of expository video: the how-to. The how-to community tends to involve a wide variety of content creators and viewers – both non-expert videographers as well as digital content creators author and access how-to videos. For this reason we felt it would be the sub-genre of expository content most representative of our prospective users.

Torrey et al. [24] interviewed authors of how-to videos to understand how their guides are generated and distributed. They found that authors utilize an array of both tools and broadcast methods to construct, iterate, and diffuse their guides. Specifically, they found that users combined video, annotated photos, text, diagrams, and other media to communicate their work. These findings suggest broadly that tools to support how-to creation and search should incorporate multimedia and mark-up. However, previous work did not investigate first-hand the methods and paraphernalia that authors actually used to create their guides.

Furthermore, while prior work [24,25] explored how electronics and computer hobbyists create how-to content, we were interested to discover the similarities and differences of how-to creation and utilization across a range of disciplines. Since our ultimate goal is to develop tools to support how-to tasks, we focused on gathering a deep, detailed understanding of tool use by sampling a diversity of contributors (see Table 1).

We interviewed 9 participants about their how-to creation practices and were able to observe two participants as they jointly created a how-to guide. While most participants were interviewed for a single hour-long session, some participants agreed to join us for additional interview sessions. Interviews and

observations were audio and/or video recorded, totaling more than 15 hours of data, and significant portions of these recordings were transcribed for analysis.

Participants were solicited largely via messaging features on how-to sharing websites such as Snapguide[1], while some were referred by contacts in the craft community. In preparation for interviews, we gathered public how-to materials posted by participants to their personal or professional websites, their blogs, and the how-to communities to which they belong. We created an interview guide and asked each participant a series of common questions, including: "How did you get started creating instructional guides", "Why do you do it?", "Do you share them? Keep them private?", "Can you think of something you did not make a how-to guide for, and why?", "What media do you use? When did video fail and why?", "What role do comments play?", and "Which hosting website do you like best for how-tos?".

We also asked each participant to step through an example, describing why they decided to make the how-to and what tools they used to produce it. However, following standard needfinding protocol, interviews were semi-structured and included many follow-up questions based on participant responses. Also, we often made use of the participants' how-to artifacts as a means to ground conversations and elicit new stories.

After collecting interview data we used an open coding scheme [23] to analyze and categorize data. Below, we share some of the implications for how-to capture tools that we drew from our distilled observations.

Capture should be unobtrusive

The recording process is often ad hoc and context-dependent, largely because authors often felt that documentation of their activity was a secondary concern. Anne, for example, said the following about her cooking guides: "If you're making [a dish] anyway, which happens to be me... I'm making it anyway, I'll take a picture of it. You know?" Derek spoke similarly about his construction projects: "I'm going to be building it, I'm just going to pick up my phone, I'm just going take a picture, going to throw the phone back down, keep going." For these participants, making and sharing how-to guides is not done for its own sake – it takes a back seat to the documented activity itself.

Participants expressed some frustration when their capture devices got in the way of the activity to be documented. For example, Anne needed both hands for shucking an oyster, so she was unable to simultaneously record video. Sometimes she, as well as Faith, Gary, and Greg, recruited friends and family to document their activity. When others were unavailable to help, Anne augmented a tripod with a rubber band to capture video, while others struggled to balance their video recorders on a sturdy surface (Derek) or simply opted not to take video at all (Elena).

In other cases, participants used tools with which they were already familiar for other reasons – such as a cell phone that can take pictures and videos –

[1] http://snapguide.com

rather than purchasing tools specifically for how-to documentation (e.g., a tripod or video editing software). Derek echoed the perspective of most other participants, saying "I would buy something if it was there for the point of construction, but I think if it was going to assist in making the [how-to], probably not." Sometimes unique circumstances led to deviations in traditional capture practices, as when Faith was on a "computerless" vacation; instead of using her cell phone to take pictures as she normally did, she used her "real" camera to take pictures for a guide.

Finally, in some cases participants used lower-quality capture and composition tools that simplified the creation process even though better equipment was at-hand. For example, though Carman and Craig owned a medium-quality DSLR camera, they still chose to use their iPhones to take pictures and videos for their food guides because of the convenience of doing everything from raw media capture to editing and sharing on the phone: "I liked the fact that it was easy and simple to upload guides, how you could do it all from one point, like all from your phone, which made it easy. I don't think I ever would have done a do-it-yourself guide if I had to take out a video camera and record it myself and then upload it to my computer."

In each of these situations, our main takeaway was that authors need to be able to choose a capture device that supports the context of the recording environment – who is available to help, what is being documented and where – so that the capture process can remain as unobtrusive as possible.

Access in chunks

Participants expressed frustration accessing how-to video content using traditional video players. Faith felt that accessing a how-to guide is, "more controlled with pictures [as compared to videos] ... YouTube videos are just too – it's just too much altogether ... I like that concept of 'this many steps.' With a video, it's just not broken down like that." Many of the other participants similarly felt that they lacked "control" over video, and that stepping through static content (e.g., photos annotated with text) was easier.

However, we doubt that this is actually a fundamental problem with video, but rather that most video tools are not designed for how-to content. Most participants did not have experience with commercial sites, such as Howcast[2], that offer interactive links into videos that can effectively help divide videos into more manageable chunks. Participants were interested in the potential of these types of interactive video tools. Elena explained that watching a how-to video while trying to enact the steps can be difficult because she often needs to replay a particular section multiple times. "If that video is 4.5 minutes and I'm not totally prepared in the right way to follow all the things, then I have to go back and start over." She "like[s] the concept of [how-to videos] being sort of chunked up for you." She also mentioned that it would be helpful to

[2] http://www.howcast.com

navigate back to a sub-section of the video to replay it or to adjust the speed of the video so that she could follow it at a natural pace.

Overall, how-to users want marks, filters, and interactive controls that are not available in typical video players.

Annotate and share

We also found that participants need to import media from other sources either to better describe a process while creating a how-to video or add other meta-data that can make the content more meaningful. Derek had difficulty "getting media into" a how-to guide and was frustrated at the overall lack of annotation capabilities and support for meta data. Other participants were frustrated by the difficulty of importing media found on the Internet that could have helped explain a particular point in their how-to.

Annotations can also be useful beyond the creation process. Anne suggested that she would like "to see a picture of [other user's] version" of a product that she documented, perhaps as an annotation on the original video, while Carman and Craig wanted to be able to construct links between segments of different how-tos.

Finally, we found that most how-to creators were motivated by the community of people watching, using, and commenting (or up-voting) their videos, but that better sharing tools would help them diffuse their work more broadly.

2.3 Mobile prototype

To gain a first-hand understanding of the kinds of issues that arise when developing a multimedia capture tool for expository content we designed a mobile prototype to help users create how-to "tips" for product operation and repair. A tip contains one or more *videos*, each of which can include any number of multimedia *bookmarks*. Bookmarks are associated with a time in the video and contain a timestamp, a keyframe, a name, and one or more *annotations*, each of which can contain any or all of: a short textual description of the bookmark; a high-resolution photo; a short audio clip, and a region marker that highlights a portion of the video frame at the time-code associated with the bookmark. In addition to its associated bookmarks, each tip can be given a name, an owner, a short description, and any number of text tags. The mobile application (implemented in Android) saves videos, bookmarks, and annotations locally to the mobile device. When the user submits a bookmark each component is serialized, transmitted to the server, and stored in a database.

Below we detail the application and describe an evaluation that revealed important issues not only for this prototype but how-to capture tools in general.

Creation

Figures 1 and 2 show the flow of events and actions for creating a new tip. The user first must associate the tip with a product using one of two options: manual entry or Optical Character Recognition (OCR) search (Figure 2a). When using OCR the mobile application opens a live camera view, runs OCR on the preview images, and sends the recognized text to the server to be matched against a database. The database may contain domain-specific details (for example a list of printer model names) or it can be a general purpose list of terms. As matching names are found they are returned and displayed to the user as clickable icons.

After associating the tip with a product name, the user can record a video. Experience with other tools suggests that bookmarking during the capture process can be a helpful way to structure video [16]. We extend support for this practice so that while recording the user can tap the screen to add a bookmark at the current time (Figure 2b). In the background, the application creates a bookmark and associates it with the current time in the video. The application also extracts a video frame from the current time and associates it with the bookmark. The user can capture all videos immediately after the scanning step and then proceed to editing, or the user can alternate between creating video captures and editing the existing set of captured videos and bookmarks.

The overview screen for captured videos and bookmarks allows the user to edit the basic details of the tip. These details include setting the name and description for the tip and optional text tags which apply to the entire tip. From this screen the user can also choose to record a new video, delete previously recorded videos, and delete or rename bookmarks. Each video capture can include a type parameter which can be used to help structure the tip. The possible capture types are “symptom”, “solution”, and “other”. The user can edit a capture by clicking on a video or any bookmark keyframe.

When editing a video capture (Figure 2c-e), a user can play the video and view or alter the contents of the associated bookmarks. New bookmarks can also be added and the positions of the existing bookmarks can be adjusted. Using the icons in the right edge of the screen, the user can add annotations to the bookmark or replace the previously added annotations. Users can swipe the screen to navigate between bookmarks. At any time, the user can press the back button to return to the overview screen. All modifications made during editing the capture are automatically saved locally to the device.

When all the necessary details are added – tip name and description, and a name for each bookmark – the tip can be uploaded (submitted) to the server. If the user did not provide all of the required information the tip cannot be uploaded and a message is shown describing what details remain to be completed.

When a tip is uploaded to the server it is automatically processed to remove unnecessary frames and extract additional metadata. If the tip author put the mobile device face down, the captured video will be completely black

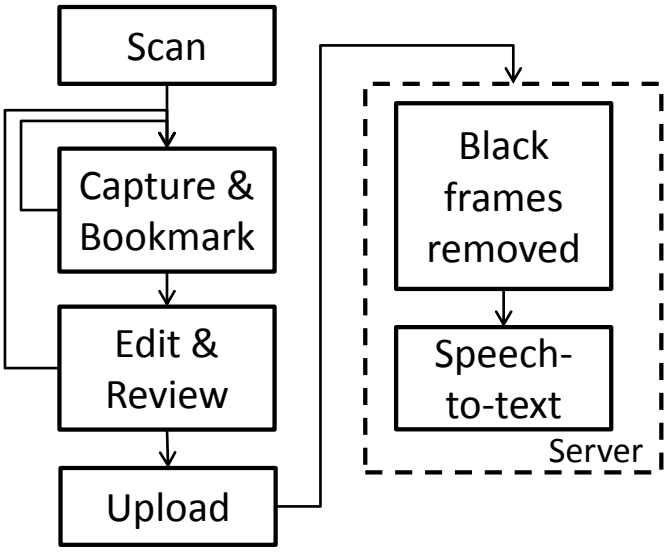


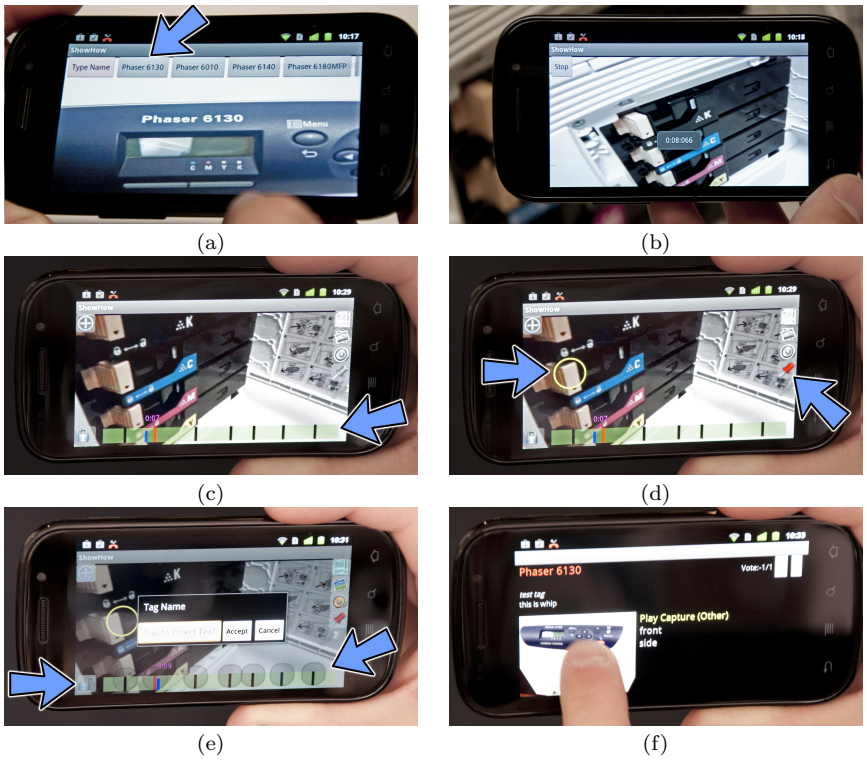
Fig. 1 Flow of tip creation activities in the mobile prototype. After scanning for (or typing) a tip title in the 'Scan' step, video captures are performed in the 'Capture & Bookmark' stage. In the 'Edit & Review' the existing video captures can be edited. Bookmarks can added, deleted, moved, or their annotations edited. Once the user is satisfied with the set of annotated video captures and bookmarks, the entire tip is uploaded to the server where the video is cleaned by removing black frames, and the audio is processed to find spoken text around bookmark points.

and is not useful for other tip users. These frames are identified and cut out of the video. Also, a tip author may utter some descriptive speech while recording a tip, in particular while marking a bookmark during capture. To help make this speech useful, off-the-shelf speech recognition algorithms (e.g., the Sphinx speech recognizer [26]) are applied to convert audio in the immediate neighborhood of a bookmark to text. If speech is recognized with a good confidence level it can be added to the searchable bookmark text to aid text-based retrieval of tips.

Search

ShowHow's search function allows users to search for tips in the database. The search system also includes a simple rating system, which allows users to vote for the quality and usefulness of the tip.

The first step, selecting a name for search, is similar to the process of selecting a name when creating a tip except that OCR results are compared only against previous tips rather than a separate product name database. Whether the user types the search term manually or uses the OCR scan to search for one, the selected term is submitted to the server and matching results are found based on word similarity between the query and the tip



(a) (b) (c) (d) (e) (f)

Fig. 2 Creating a tip. (a - Scan) Before shooting the video, live frames are sent to an OCR engine and then on to the server to determine the model and make of the product. The most likely model names appear as clickable icons at the top of the screen (arrow). (b - Capture & Bookmark) While recording, users can tap the screen to set a bookmark. (c - Edit & Review) While editing, the video bookmarks appear as black bars on the video timeline (arrow). Selecting a bookmark turns its bar red. (The interface for viewing a tip is similar but with editing functionality disabled.) (d) Users can add annotations to each bookmark. Here the user adds a circular region tag (left arrow). When a new annotation type is added its icon is illuminated (right arrow). (e) The user can unlock the timeline by pressing the lock icon (left arrow). The circle around each bar (right arrow) indicates its touch area. Pressing and moving a bar changes the bookmark's position in the timeline. Pressing and holding a bar lets the user set the name of its associated bookmark. On the overview screen (not shown) the user can associate tags and a more detailed text description with the entire tip and then upload all of the content to the server. (f) Immediately after uploading a video it can appear in search results.

names. A list of tips is returned to the user showing a preview image and summary of each tip (Figure 2f). The interface also allows users to see how others have rated (voted on) the tip and provides a button to vote the tip up or down.

Evaluation

We ran a laboratory study focused on tip creation, since it is far more elaborate than tip search. The study was composed roughly of two parts. First, we asked participants to create a tip following step-by-step instructions in order to become minimally familiar with the tool. We did not save any of this data for analysis. Second, we asked participants to use the tool to create a tip describing how to setup some components of a computer. This involved connecting a printer and a webcam to the computer, using the webcam application on the desktop to take a picture to verify that it is working properly, and printing a picture using the webcam application to verify that the printer was also working properly.

After running two pilot users to iterate the study parameters another six participants (two women) completed the study. Since other mobile video creation tools focus on constructing a narrative rather than a browseable, searchable tip, we focused on verifying the feasibility of our approach rather than comparison with other systems. For this reason we did not record timing information, but rather asked users to openly discuss issues they had as they completed the task. We also asked participants to rate the usefulness of each annotation type at the end of the study.

Our findings revealed a tension between simplicity and expressiveness in a mobile creation tool. While all participants but one completed the task, many expressed interest in features closer to traditional video editing, such as audio track replacement or rich composition features. One participant noted that adding bookmark annotations can make changes to the narration necessary. Specifically, he narrated one section of one of his videos, "well, you can't really make this out, but..." Later, though, after adding a high resolution photo that helped clarify what he was referring to, he wanted to be able to override the audio track (not only add an annotation) by saying instead, "look at the image in the [bookmark] to see this more clearly." Two other participants were similarly happy with the visuals they had shot but wanted to rehearse and record an entirely new audio track. Also, one participant made use of the black screen removal feature to insert pauses into the video by placing the camera down on the table while it was still recording. But another two participants wanted to remove content at the beginning and end of videos post hoc.

Overall, text, images, and audio were the most frequently used and most valued annotations. Half of the participants found Android's speech-to-text functionality useful for filling in text fields. The region tag seemed to cause some confusion, perhaps because the screen was already overloaded with so many touch options. Participants were split on the way they created tags, with half creating them while shooting and the other half creating them from the timeline. Also, two participants wanted much more control over the video while shooting. One participant wanted to, "touch the screen to meter and focus on a region of the screen." This is a potentially useful feature since touching the

screen could also bootstrap the location of the region tag, ameliorating some of the issues participants had placing it post hoc.

Finally, many participants also had difficulty holding the device while trying to also position the object being recorded. While tripods for mobile devices could address this issue (e.g., Joby[3]), one participant also suggested using a head-mounted camera (e.g., Looxcie[4]).

Related work

Recent work has investigated the use of systems to connect experts to users in the field in real time [17]. Our approach shares more in common with systems such as View[5] and Howcast in which users create and share tips with one another asynchronously. However, unlike these systems, the ShowHow capture client supports recording videos with a mobile device. Furthermore, this application improves the usefulness of recorded content by helping users add annotations and other meta-data to their videos at the point of capture. While past work has shared this goal (e.g., [13]), such systems tend to focus on adding contextual data external to the video rather than augmenting the video content itself. Other systems (e.g., [28]) support editing but do not facilitate annotation.

2.4 Summary

Torrey et al. found that how-to "sharing occurs within and across a collection of communication tools without any centralized control" [24] and that people tended to find information by browsing as much as by more directed searches [25]. We found that tools for the capture, creation, and access of how-to guides were similarly decentralized. This finding led us to develop tools to capture and convert content created in different formats to a similar view. Content creators and learners alike can use a single web-based application to browse, skim, and view content from a variety of sources.

3 Web-based tools for ingest, annotation, and access

We built a collection of tools that can support the more decentralized approach to content creation revealed by our needfinding studies:

Capture
: In addition to the ShowHow mobile client described above, users can upload arbitrary videos to our server or specify a URL of a YouTube video to annotate. We have also built tools to convert different types of how-to content into annotateable videos (see *Lightweight content* below).

[3] http://joby.com/store/gorillamobile
[4] http://www.looxcie.com
[5] http://view.io

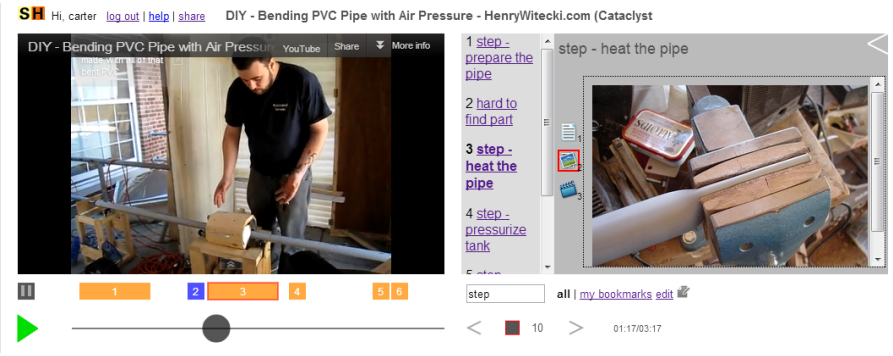


Fig. 3 The main ShowHow web viewer and editor. Numbered rectangles just above the timeline are bookmarks, and their associated multimedia annotations are at the upper-right. Users can filter bookmarks using the live search box (here, all of the bookmarks with "step" in the title are highlighted) and can limit the bookmarks shown to only those they have created ("my bookmarks"). The interface also allows them to skim between bookmarks or 10-second chunks of video. Finally, users can share whole videos or segments with the "share" button.

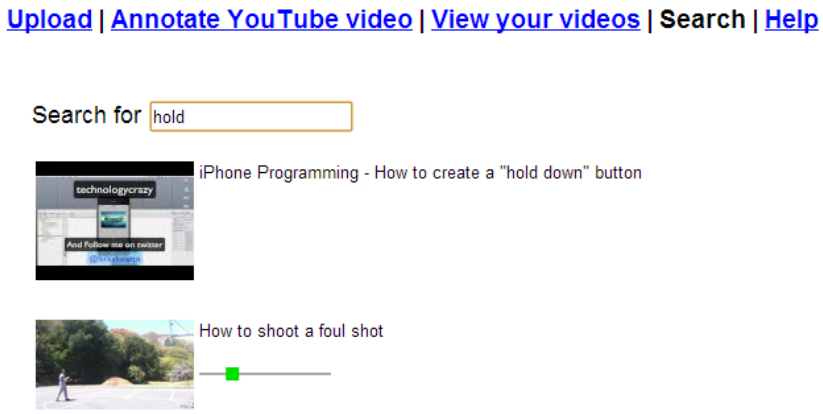


Fig. 4 The ShowHow web search tool. Here the search text has matched a video's title (top) as well as a specific bookmark within a video (bottom) as indicated by the green mark on the timeline indicating the point of the match.

View and annotate

Inspired by other tools that use bookmarks to index interface actions [2] and other events[6], we built an HTML5-based web client to support creation, editing, sharing, and viewing of bookmarked, annotated videos (see Figure 3). This client is designed for desktops and tablets and supports all of the bookmarking and annotation functionality of the mobile client. Users can also filter bookmarks by content-creator or with a live search of the bookmark's title and text annotations.

[6] http://teachscape.com

Mobile viewer
: We also built a separate viewer explicitly for mobile phones. This simplified interface allows users to swipe the screen to navigate between bookmarks (or 10-second chunks if there are no bookmarks present). The main web viewer will automatically switch to this view if it detects a small-screen device.

Search
: Videos and their associated bookmarks and annotations are searchable in a separate interface. Search results can link to an entire video or a specific bookmark within a video (see Figure 4).

We can use these tools to view a range of expository video types.

3.1 Lightweight content

For lightweight expository content, such as guides or how-to videos, bookmarks are typically either explicit steps or important points within the video. Since the bookmarks in ShowHow have no inherent semantic meaning, they can be used flexibly. The bookmarks in Figure 3 are a combination of steps (bookmarks 1 and 3-6) as well as one that marks an important point (bookmark 2).

The ShowHow interface supports manually adding bookmarks to any type of video. For certain types of videos, important points or steps can be extracted automatically. As an example consider another mobile multimedia tool we built, SketchScan[7], that allows the user to capture an image, select regions in the image and associate audio annotations with them, and finally generate a video from the sequence of image regions and their audio annotations (Figure 6). We built the SketchScan system before ShowHow, but the content it yields lends itself easily to bookmarking. Without changing the SketchScan code at all, we built an ingest tool for ShowHow that can detect boundaries in SketchScan videos by correlating breaks in the audio with global changes in the video's image content.

We built a similar ingestion tool for the Snapguide system. Snapguide is a third-party application that makes it easy to create and publish multimedia instructions with a mobile tool. It differs from the ShowHow mobile application in that it is fundamentally step-based – users create steps first and then fill them in with content, such as a photo, video clip, and text. The interface that SnapGuide uses to display individual guides therefore more closely resembles traditional step-by-step recipe guides than a how-to video. Inspired by work that combines static and dynamic content in online tutorials [5], we were able to create an ingest tool to ShowHow that converts each step, which may be a video clip or a photo and may include optional text, into a single video with text-based annotations (see Figure 5). This conversion process allows us to view yet another type of how-to content in a consistent interface.

[7] http://sketchscan.fxpal.com

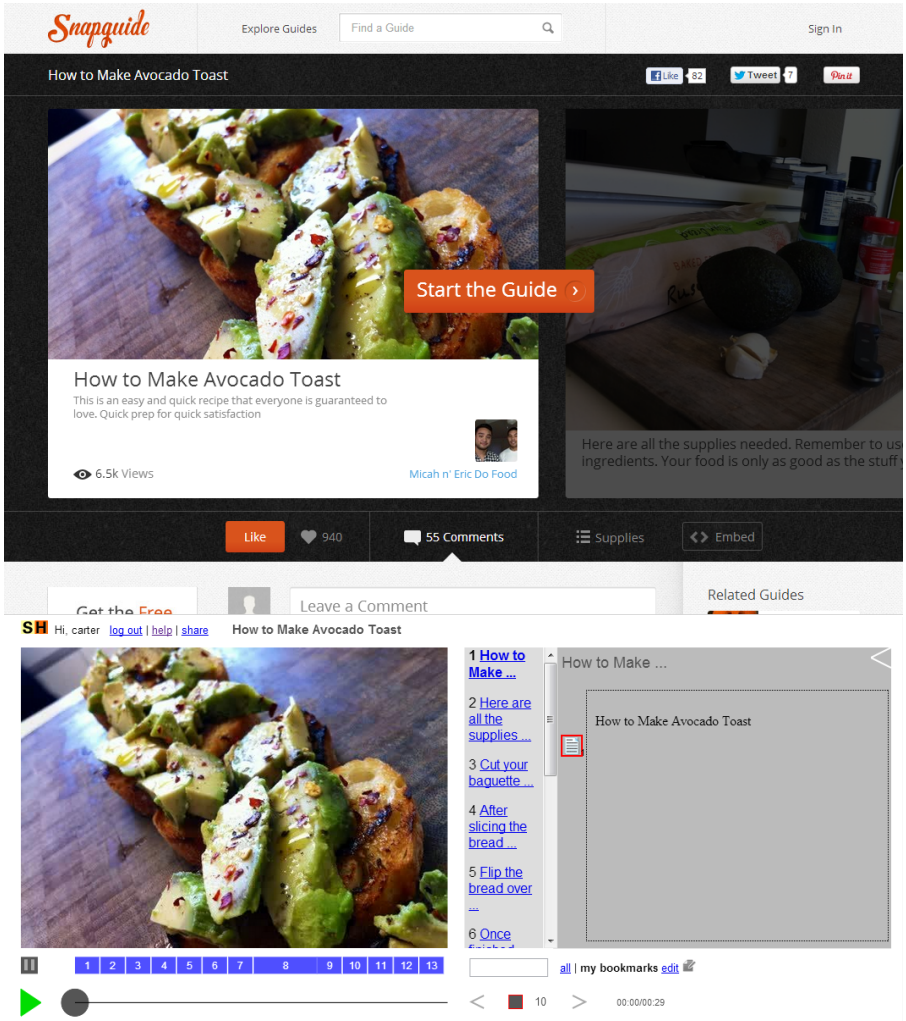


Fig. 5 Converting from Snapguide to ShowHow. How-to content in Snapguide is presented in a step-based viewer (top) but can easily be converted to ShowHow's video-and-bookmark-based approach (bottom).

3.2 Produced content

For content that was produced more deliberately, bookmarks tend to have more explicit mappings, such as title screens or, in lecture videos, transitions between slides.

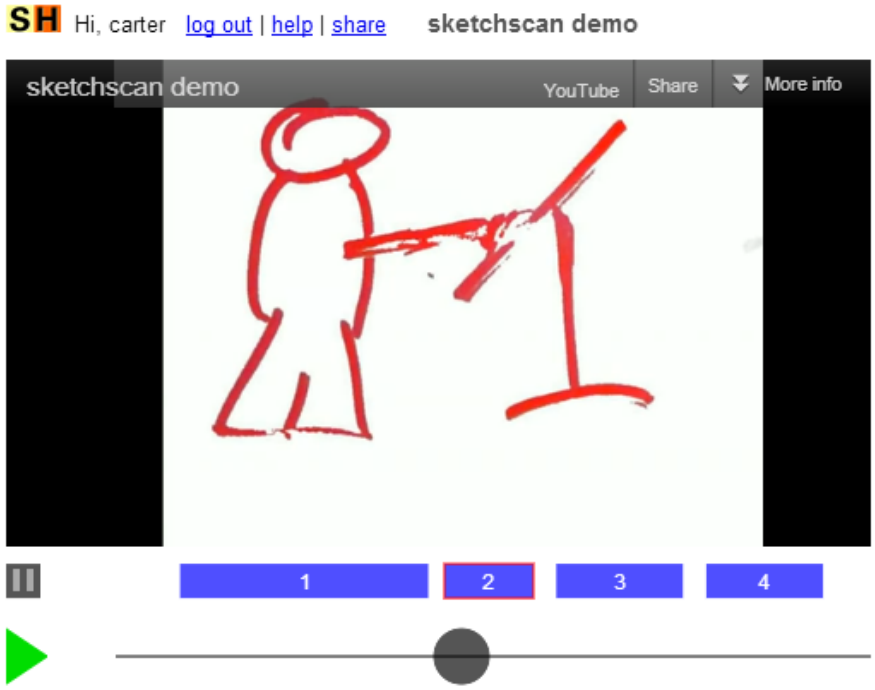


Fig. 6 Viewing a SketchScan video in ShowHow. SketchScan produces videos from audio annotations attached to regions of a captured image. A ShowHow ingest tool determined that this video contained four different audio annotations and added corresponding bookmarks to the video, shown here in the ShowHow player.

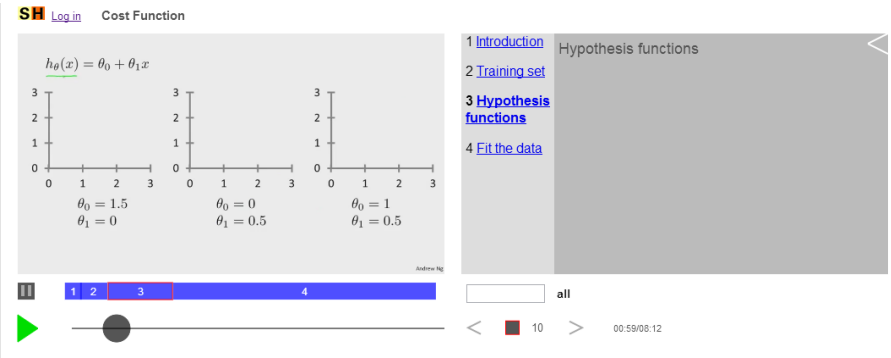


Fig. 7 ShowHow can support online lecture content. Here, bookmarks represent slide transitions that were detected automatically. Bookmark titles were extracted manually from the video's transcript.

Our experiments in this domain have thus far focused on lecture content available through Coursera[8], which recently enrolled its one millionth student[9], and specifically the Machine Learning course[10]. The course's video content is a blend of shots of the instructor speaking and shots of a slide stream with audio commentary. The videos also include a substantial amount of handwritten annotation with electronic ink overlaid on the slides.

To ingest these videos into ShowHow, the primary goal is to automatically temporally segment the videos according to the slides that are shown and discussed. The slides typically reflect the presenter's topical structuring of the content. We aim to leverage this structuring to facilitate video browsing and navigation via bookmarks in ShowHow. Detected slide segments are associated with corresponding bookmarks that pre-populate the ShowHow player. Bookmarks can then be deleted, added, or augmented with annotations by users.

The video analysis includes three components. The first is a support vector machine (SVM) classifier which discriminates shots of the presenter from shots of slides. This classifier was trained on standard (RGB) color histogram features computed over a non-uniform spatial grid that emphasizes the center of the frame. The classifier was trained using video frames from an optional tutorial section of the course that were manually labeled. Classification was found to be more reliable than face detection which exhibited a high rate of false positives among the electronic ink annotations.

The second component is a simple frame difference detector adapted from previous work on lecture video analysis [1]. We sample video frames every second and compare temporally adjacent frames pixel-wise. We filter out regions of change with low spatial support, and sum the remaining number of pixels above threshold. When the number of changed pixels exceeds 45% of the video frame area, we detect a slide change. This simple approach has been reliable in our initial experiments. The final component is a more refined frame difference analysis designed to detect the addition of electronic ink annotations. As before, we apply spatial filtering to remove small changed pixel regions. We then sum the number of remaining changed pixels and declare a new annotation if the sum exceeds 10% of the frame area. In this case, we also apply a stability constraint. Specifically, we save a keyframe that includes the new annotation after the number of changed pixels in the inter-frame difference image remains below the 10% threshold for at least two seconds. This is done to avoid detecting multiple incomplete versions of a single annotation.

Given a source video, we first apply the SVM classifier to detect frames that contain the speaker. We next sample the video one frame per second and compute the inter-frame differences as above. From this processing, we construct a two level temporal segmentation. The first level includes each unique slide-based segment. In the second level, we include the times at which

[8] http://coursera.com

[9] http://blog.coursera.org/post/29062736760/coursera-hits-1-million-students-across-196-countries

[10] https://www.coursera.org/course/ml

any complete ink annotations are overlaid on each slide. Shots of the speaker are currently not included in this segmentation, since they usually provide little useful visual context to aid in video navigation.

Figure 7 shows an example in which the slide-based segments appear as bookmarks for one of the Coursera videos in our corpus. Currently the user can enter bookmark titles manually – we also plan to develop tools to derive them automatically or semi-automatically using OCR and transcript data.

4 Conclusion and future work

Past systems for disseminating how-to knowledge focused on textual descriptions of problems and solutions [27]. In contrast, our goal with this work is to explore the use of multimedia to both capture and represent tacit information as well as relevant contextual cues.

The tools we developed for this purpose are relevant for how-to descriptions as well as a wider range of expository multimedia content. A series of case studies as well as a mobile prototype led us to the conclusion that the best tools are those that flexibly incorporate a variety of other tools that themselves support different styles of capture and access. This led directly to the design of an HTML5-based video annotation system that supports automated bookmark generation for some content as well as manually added bookmarks and multimedia annotations.

This new tool was designed to work across a variety of platforms including desktops, tablets, and phones. The next steps for the work are to deploy the new tools more broadly to better understand their ability to support both lightweight how-to content as well as more professionally produced content.

Furthermore, past work has suggested that first-person video instructions can improve performance on assembly [12] and learning [14] tasks. We are currently investigating methods to better integrate head-mounted capture systems in order to generate interactive video tutorials from the user's viewpoint.

Finally, one consistent finding is that static visuals, such as still images and diagrams, and dynamic visuals, such as animations and videos, support different types of learning. Specifically, static visuals "promote understanding of processes," while animated visuals better convey procedures [6]. We are currently extending our HTML5 tool suite to seamlessly weave together a variety of media types, including static images, animated images, videos, and audio. This will allow content creators complete flexibility in conveying both how to complete hands-on tasks as well as the fundamental processes underlying them.

References

1. Adcock, J., Cooper, M., Denoue, L., Pirsiavash, H. and Rowe, L. A. Talkminer: A lecture webcast search engine. ACM MM. 241–250. 2010.
2. Banovic, N., Grossman, T., Matejka, J. and Fitzmaurice, G. Waken: Reverse engineering usage information and interface structure from software videos. ACM UIST. 83–92, 2012.
3. Barthel, R., Ainsworth, S. and Sharples, M. Collaborative knowledge building with shared video representations. International Journal of Human Computer Studies. 71(1). 59–75. 2013.
4. Beyer, H. and Holtzblatt, K. Contextual design: Dening customer-centered systems. Series in Interactive Technologies. Morgan Kaufmann, San Francisco. 1998.
5. Chi, P-Y., Ahn, S., Ren, A., Dontcheva, M., Li, W. and Hartmann, B. MixT: Automatic generation of step-by-step mixed media tutorials. ACM UIST. 93–102. 2012.
6. Clark, R. C. and Mayer, R. E. E-Learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning. Pfeiffer, San Francisco. 2011.
7. Eiriksdottir, E. and Catrambone, R. Procedural instructions, principles, and examples: How to structure instructions for procedural tasks to enhance performance, learning, and transfer. Human Factors. 53(6). 749–770. 2011.
8. Grossman, T. and Fitzmaurice, G. ToolClips: An investigation of contextual video assistance for functionality understanding. ACM CHI. 1515–1524. 2010.
9. Hammersley, M. and Atkinson, P. Ethnography: Principles in Practice. Routledge, London. 1995.
10. Harrison, S. M. A comparison of still, animated, or nonillustrated on-line help with written or spoken instructions in a graphical user interface. ACM CHI. 82–89. 1995.
11. Holtzblatt, K., Wendell, J. B. and Wood, S. Rapid contextual design: A how-to guide to key techniques for user-centered design. Morgan Kaufmann, San Francisco. 2005.
12. Kraut, R. E., Fussell, S. R. and Siegel, J. Visual information as a conversational resource in collaborative physical tasks. Human-Computer Interaction. 18(1). 13–49. 2003.
13. Lahti, J., Westermann, U., Palola, M., Peltola, J. and Vildjiounaite, E. MobiCon: Integrated capture, annotation, and sharing of video clips with mobile phones. ACM MM. 798–799. 2005.
14. Lindgren, R. Generating a learning stance through perspective-taking in a virtual environment. Computers in Human Behavior. 28(4). 1130–1139. 2012.
15. Moreno, R. and Ortegano-Layne, L. Do classroom exemplars promote the application of principles in teacher education? A comparison of videos, animations, and narratives. Educational Technology Research and Development. 56(4). 449–465. 2008.
16. Niu, J., Huo, D., Xie, X., Lin, J., Zeng, X. and Liu, Y. MoViShooter: Bookmarking videos in real-time for mobile device users. ACM CHI Workshop on Video interaction. 2011.
17. O'Neill, J., Castellani, S., Roulland, F., Hairon, N., Juliano, C. and Dai, L. From ethnographic study to mixed reality: A remote collaborative troubleshooting system. ACM CSCW. 225–234. 2011.
18. Palmiter, S., Elkerton, J. and Baggett, P. Animated demonstrations vs. written instructions for learning procedural tasks: A preliminary investigation. International Journal of Man-Machine Studies. 34(5). 687–701. 1991.
19. Patnaik, D. Needfinding: Design research and planning. CreateSpace Independent Publishing Platform. 2013.
20. Pirolli, P. Effects of examples and their explanations in a lesson on recursion: A production system analysis. Cognition and Instruction. 8(3). 207–259. 1991.
21. Pongnumkul, S., Dontcheva, M., Li, W., Wang, J., Bourdev, L., Avidan, S. and Cohen, M. F. Pause-and-play: Automatically linking screencast video tutorials with applications. ACM UIST. 135–144. 2011.
22. Schwan, S. and Riempp, R. The cognitive benets of interactive videos: Learning to tie nautical knots. Learning and Instruction. 14(3). 293–305. 2004.
23. Strauss, A. and Corbin, J. Basics of qualitative research: Grounded theory procedures and techniques. Sage, Newbury Park, CA. 1990.
24. Torrey, C., McDonald, D., Schilit, W. and Bly, S. HowTo pages: Informal systems of expertise sharing. ECSCW. 391–410. 2007.

25. Torrey, C., Churchill, E. F. and McDonald, D. W. Learning how: The search for craft knowledge on the internet. ACM CHI. 1371–1380. 2009.
26. Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P. and Woelfel, J. Cambridge, Sphinx-4: A flexible open source framework for speech recognition. Technical Report. 2004.
27. Whalen, J. and Bobrow, D. G. Communal knowledge sharing: The Eureka story. Chapter in Making work visible: Ethnographically grounded case studies of work practice, edited by Margaret H. Szymanski and Jack Whalen. Cambridge, UK: Cambridge University Press. 257–284. 2011.
28. Wu, C-I., Teng, C. J., Chen, Y-C., Lin, T-Y., Chu, H-H. and Hsu, J. Y. Point-of-capture archiving and editing of personal experiences from a mobile device. Personal and Ubiquitous Computing. 11(4). 235–249. 2007.
29. Zahn, C., Pea, R., Hesse, F.W. and Rosen, J. Comparing simple and advanced video tools as supports for complex collaborative design processes. Journal of the Learning Sciences. 19(3). 403–440. 2010.
30. Zhang, D., Zhou, L., Briggs, R. O. and Nunamaker Jr, J. F. Instructional video in e-learning: Assessing the impact of interactive video on learning effectiveness. Information & Management 43(1). 15–27. 2006.