Supporting media bricoleurs*

Scott Carter, Matthew Cooper, Laurent Denoue, John Doherty, Vikash Rugoobur
FX Palo Alto Laboratory
{carter,cooper,denoue,doherty,vik}@fxpal.com

Online video is incredibly rich. A 15-minute home improvement YouTube tutorial might include 1500 words of narration, 100 or more significant keyframes showing a visual change from multiple perspectives, several animated objects, references to other examples, a tool list, comments from viewers and a host of other metadata. Furthermore, video accounts for 90% of worldwide Internet traffic[1]. For new startups, it has become de rigueur to explain their new products with video rather than text and still graphics. This is likely because more people spend time consuming video than text. For example, the NEA reports that Americans aged 15-24 watch TV two hours per day but read for leisure only seven minutes per day [NEA]. YouTube reports that "[o]ver 6 billion hours of video are watched each month on YouTube — that's almost an hour for every person on Earth, and 50% more than last year [2011]... 100 hours of video are uploaded to YouTube every minute"[2]. Video is undeniably an increasingly prominent consumer communication medium.

However, it is our observation that video is not widely seen as a full-fledged document; dismissed as a media that, at worst, gilds over substance and, at best, simply augments text-based communications. "Idiot box" and "boob tube" are listed as synonyms with TV in the Merriam-Webster dictionary. Even educational videos found in MOOCs have been derided as "unsophisticated chunks" [Vardi]. But there is no overwhelming evidence that static media better conveys knowledge or engenders higher quality thinking than temporal media. In some cases increased television viewing correlates positively with IQ[3]. Indeed, humans "were never born to read" — visual storytelling predates the written word by thousands of years [Wolf]. In this piece, we suggest that negative attitudes toward multimedia documents that include audio and video are largely unfounded and arise mostly because we lack the necessary tools to treat video content as first-order media or to support seamlessly mixing media.

Building video-based interfaces is challenging. One difficulty is the "semantic gap" that characterizes machine representations of non-textual media [Hauptmann]. Words are natural compositional features representing text documents, and are endowed with objective meanings. In contrast, prevalent feature representations of visual and auditory content (e.g., SIFT, MFCC, etc. [Ye, et al.]) are neither grounded by meaning nor provide a natural structuring for manipulation or reuse. As a result automatic tools for decomposing multimedia content into coherent subunits or characterizing their semantics are relatively primitive in comparison to text.

Additionally, many existing tools treat video monolithically, rather than as a potentially interactive, mineable, sharable, and reconfigurable medium. Many startup systems exist that allow users to remix video, but they tend to operate breadth-first; simply allowing users to string

[1] http://technews.tmcnet.com/iptv/topics/iptv/articles/136034-video-dominates-global-traffic.htm
[2] http://www.youtube.com/yt/press/statistics.html
[3] http://www.psychologytoday.com/blog/the-human-beast/200903/does-watching-tv-make-us-stupid/

together clips rather than organizing or exposing the content buried within. Research has focused on the related problem of understanding and developing visual literacy toward the production of video ([e.g., [Weilenmann]). While this work is valuable it keeps media locked into a particular representation. In his book *Mindstorms*, Papert suggests that "in the most fundamental sense, we, as learners, are all bricoleurs" and that we build our understanding of complicated processes by tinkering and reconfiguration [Papert]. But in order to tinker you need building blocks and current tools do little to expose those tinkerable components from this ubiquitous medium.

While text is the ideal way to express many ideas, a great deal of text content is created to document multimedia (visual, aural, or multi-modal) information: descriptions or reviews of music, paintings or other art; processes or demonstrations of a technique; system construction or deconstruction, etc. However, marrying abstract analysis in one medium with a medium verisimilitudinous with the described content facilitates a broader understanding of pragmatic concepts. For example, just as we might expect a preview audio clip to accompany a new album's review, it is no doubt helpful to include video alongside text that is fundamentally procedural (e.g., JoVE[4]), or interactive demonstrations alongside descriptions of computational concepts (e.g., Learnable Programming[5]). Furthermore, there is work showing that gestures are at least as important (if not more so) than the spoken words they complement [Mehrabian]. Video and other animated content can similarly complement text. As previous work has shown, when expository video is made interactive it can be useful above and beyond in-person lectures because video can be repeated, accessed randomly, [Zhang] and annotated [Zahn].

But we would like to go a step further than interaction and see video content that is fully ready-to-hand [Heidegger] so that learners and creators can engage in the bricoleur's "dialogue with the materials" [Lévi-Strauss]. Once video content can be manipulated using the same techniques and metaphors we apply to text such as cut-and-paste, drag-and-drop, and spatial editing, we can build tools that support the construction of multimedia documents that richly convey procedural and analytical content in concert with the most appropriate media.

We further suggest that we need tools that focus on content rather than markup. When he created HTML, Tim Berners-Lee never intended for people to "have to deal with HTML" [Berners-Lee]. Multimedia documents have been supported somewhat (e.g. wikis), but these tools are not conceptually different from HTML — they still require users to markup text rather than directly manipulate content.

Media bricolage tools must allow users to extract media so that it can be seamlessly remixed in multimedia documents. But what exactly do we mean by "multimedia document"? For our purposes, a multimedia document does not simply place different pieces of multimedia in proximity — web sites have done this quite well for years. Rather, we mean documents in which spatial and temporal layouts have equal weight, can influence one another, and through which content can flow in any direction. Traditional documents are designed to be consumed spatially,

[4] http://www.jove.com/
[5] http://worrydream.com/LearnableProgramming/

while videos are designed for temporal navigation. In a multimedia document the goal is to take advantage of a traditional document's spatial qualities to augment video and vice versa. Spatial navigation events should be able to trigger changes in primarily time-based media. Several web-based journalism sites have been exploring this approach[6]. For example, in ESPN's longform piece on the Iditarod[7], the reader follows the author as he travels through Alaska across the course. As the reader scrolls, a map at the top tracks his progress. Similarly, in the New York Times piece Snow Fall, animations respond to a user's spatial navigation[8]. Multimedia documents should also support spatial changes triggered by temporal events. For example, Mozilla's *PopcornMaker* tool allows content creators to trigger the appearance of documents when a video reaches a certain timepoint (SMIL-based authoring tools have supported similar features for many years)[9].

We can expand the idea of responsive documents more broadly to also include spatial events that trigger other spatial changes (e.g., a background changes as the user navigates) and temporal events triggering other temporal changes (e.g., pausing a video upon reaching a marked time and then playing an animated GIF in a separate window to emphasize a point).

It is important that content easily flows between media types so it can be tightly integrated across media. In our lab we are currently developing a suite of tools to support such seamless inter-media synthesis. The suite, called *Cemint* (for Component Extraction from Media for Interaction, Navigation, and Transformation), includes mobile- and web-based tools that allow users to create temporal content from spatial resources and vice versa. *SketchScan*, a mobile application we built that lets users capture, clean, animate, and share sketches, is a demonstration of the former[10]. With this app, users define regions of a sketch, optionally add audio annotations to each region, and a cloud service generates a movie from the sketch and annotated regions (Figure 1). In *SketchScan* users do not actually shoot video. Instead, the system creates a video from a sequence of multimedia bookmarks. Each bookmark includes a highlighted subregion of an image and an optional audio clip. Users capture a static image then create bookmarks and arrange them to tell a story. The sequenced bookmarks and their annotations are then forwarded to a remote server that combines them all into a single video.

The reverse case, extracting media from videos for use in static documents, has been explored previously mostly for summarization purposes. For example, video summary tools have been developed that extract keyframes into a pleasing static design [Uchihashi]. But there are many other ways to take advantage of video content in user interfaces. As part of our *Cemint* suite we are building tools that allow users to extract any keyframe from a video, or automatically-detected sub-regions of keyframes, at any time. With these tools users directly interact with video content using familiar techniques such as dragging a selection box over an area to highlight text, mouse-wheel to scroll up and down, or double-click to identify rectangular areas

[6] http://interactivenarratives.org/
[7] http://www.grantland.com/story/_/id/9175394/out-great-alone/
[8] http://www.nytimes.com/projects/2012/snow-fall/
[9] https://popcorn.webmaker.org/
[10] http://sketchscan.fxpal.com/

of importance (Figure 2). Users can then use familiar copy-and-paste or drag-and-drop techniques to extract content to multimedia documents [Denoue].

One side effect of supporting flexible repurposing of content across media is allowing users to construct thoughts in the best domain for ideation. Users can then reuse media directly without having to shoehorn their work to fit a particular tool. This could be a boon for new learners — as a previous *Interactions* paper points out, many novice users of word processors tend to spend more time constructing their thoughts outside the context of the program than within the word processor itself [Huh].

A second side effect is a reconsideration of the interplay between automatic multimedia analysis and user interaction. Content-based analysis has been dominated by a fully automatic, end-to-end paradigm in which the system takes content as input and produces a semantic representation or structuring as output. While tremendous progress has been made in content-based processing, robustness has remained elusive in part because the tasks for which these systems are designed are 'disconnected from [their] application' [Worring]. In our present work, we envision deploying much more flexible analysis tools that leverage user input as a key component of the analysis. User input is solicited using natural, familiar interactions with both the content and exposed intermediate results of real time analysis. This view has the potential to mitigate many existing obstacles to deploying automatic analysis in consumer applications.

We are just beginning our work in this area — we are far from providing fully fledged multimedia document support. There are many other ways to apply document concepts to help users navigate and extract content from video. For example, we are currently exploring how real-time analysis of live video, such as video conferences or lectures, can enable better notetaking, review, and content reuse. Tools or techniques that make it easy for spatial navigation to trigger side effects that enrich the reading experience without distracting from main concepts represents another gap in current support. Finally, we believe that better integration of video and demonstration tools could dramatically improve the way that many research results in the HCI community are communicated.

“If your medium doesn't easily allow you to correct mistakes, knowledge will tend to be carefully vetted. If it's expensive to publish, then you will create mechanisms that winnow out contenders. If you're publishing on paper, you will create centralized locations where you amass books... Traditional knowledge has been an accident of paper.” [Weinberger]

The main goal for any multimedia document tool is to allow users to tell a story using the most appropriate combination of rich and traditional media. As reading continues to move to mobile and tablet devices, a rich, multimedia approach will increasingly be the most natural way to convey formal and informal concepts. Ultimately, this will lead to a reformulation of the very notion of “knowledge.”

*Bricoleur: one who participates in “bricolage,” construction (as of a sculpture or a structure of ideas) achieved by using whatever comes to hand

References

Tim Berners-Lee. On Simplicity, Standards, and "Intercreativity". WWW Journal. 1(3). 3-10. 1996.

Laurent Denoue, Scott Carter, and Matthew Cooper. 2013. Content-based copy and paste from video documents. In Proceedings of the ACM symposium on Document engineering. 215-218. 2013.

Alexander Hauptmann, Rong Yan, Wei-Hao Lin, Michael Christel, and Howard Wactlar. Can High-Level Concepts Fill the Semantic Gap in Video Retrieval? A Case Study With Broadcast News. IEEE Transactions on Multimedia. 9(5). 958-966. 2007.

Martin Heidegger. Being and Time. Harper and Row, 1962.

Joohee Huh. Why Microsoft Word does not work for novice writers. interactions 20(2). 58-61. 2013.

Claude Lévi-Strauss. The Savage Mind. University of Chicago Press. 1966.

Albert Mehrabian. Silent Messages: Implicit Communication of Emotions and Attitudes. Wadsworth. 1972.

NEA. To Read or Not To Read: A Question of National Consequence. 2007.

Seymour Papert. Mindstorms: Children, Computers, and Powerful Ideas. Basic Books. 1993.

Shingo Uchihashi, Jonathan Foote, Andreas Girgensohn, and John Boreczky. Video Manga: generating semantically meaningful video summaries. In Proceedings of the ACM international conference on Multimedia. 383-392. 1999.

Guangnan Ye, I-Hong Jhuo, Dong Liu, Yu-Gang Jiang, D. T. Lee, and Shih-Fu Chang. Joint audio-visual bi-modal codewords for video event detection. In Proceedings of the ACM International Conference on Multimedia Retrieval. Article 39. 2012.

Moshe Y. Vardi. Will MOOCs destroy academia? Communications of the ACM 55(11). 5. 2012.

David Weinberger. Too Big to Know: Rethinking Knowledge Now That the Facts Aren't the Facts, Experts Are Everywhere, and the Smartest Person in the Room Is the Room. Basic Books. 2012.

Alexandra Weilenmann, Roger Säljö,Arvid Engström. Mobile video literacy: Negotiating the use of a new visual technology. Personal and Ubiquitous Computing. http://dx.doi.org/10.1007/s00779-013-0703-x.

Maryanne Wolf. Proust and the Squid: The Story and Science of the Reading Brain. Harper Perennial. 2008.

Marcel Worring, Paul Sajda, Simone Santini, David A. Shamma, Alan F. Smeaton, and Qiang Yang. 2012. Where Is the User in Multimedia Retrieval?. IEEE MultiMedia 19(4). 6-10. 2012.

Carmen Zahn, Roy Pea, Friedrich W. Hesse, and Joe Rosen. Comparing simple and advanced video tools as supports for complex collaborative design processes. Journal of the Learning Sciences. 19(3). 403-440. 2010.

Dongsong Zhang, Lina Zhou, Robert O. Briggs, and Jay F. Nunamaker Jr. Instructional video in e-learning: Assessing the impact of interactive video on learning effectiveness. Information & Management 43(1). 15-27. 2006.

Figure captions

Figure 1. SketchScan overview screen with the second of three bookmarks selected (a). The bookmark includes a region of a static image as well as an audio clip. Users can rearrange the order of clips (b). When users are satisfied with their bookmarks and annotations they send the data to a server, which generates a video.

Figure 2. Directly interacting with video content with Cemint. Users can highlight text (a & b); manipulate the mouse wheel to scroll (c & d); and select regions of importance (e & f).

Biographies

Scott Carter is a senior research scientist at FX Palo Alto Laboratory. His primary research focus is developing innovative multimedia user interfaces.

Matthew Cooper is a senior research scientist at FX Palo Alto Laboratory, leading the Interactive Media group. His primary research focus is developing content analysis techniques that enable multimedia information management and retrieval applications. He is a senior member of both the IEEE and the ACM.

Laurent Denoue is a researcher at FX Palo Alto Laboratory interested in user interaction design, document and video processing. Laurent worked on XLibris, an annotation system; ProjectorBox, an appliance for capturing meetings; TalkMiner, a service that detects slides in

online lectures. His recent interests are client-based video processing to manipulate video documents in real-time.

John Doherty is a Senior Media Specialist at FX Palo Alto Laboratory. His primary interest is in designing processes and systems that make video easier to produce, repurpose and integrate into multimedia documents. His collaborations include: NudgeCam media capture, the mBase video indexing system and the Hitchcock Semi-Automatic Video Editor.

Vikash Rugoobur is a visiting researcher at FX Palo Alto Laboratory. His interests include robotics, quantified self, wearable devices and user interaction. Vikash co-created Sonny, a telemedicine rehabilitation platform for children with TBI; EDay, a prototype tablet controlled and connected electric vehicle; and a Microsoft Kinect instrumented drivable prototype car, capable of capturing various driver metrics and providing gesture functionality.

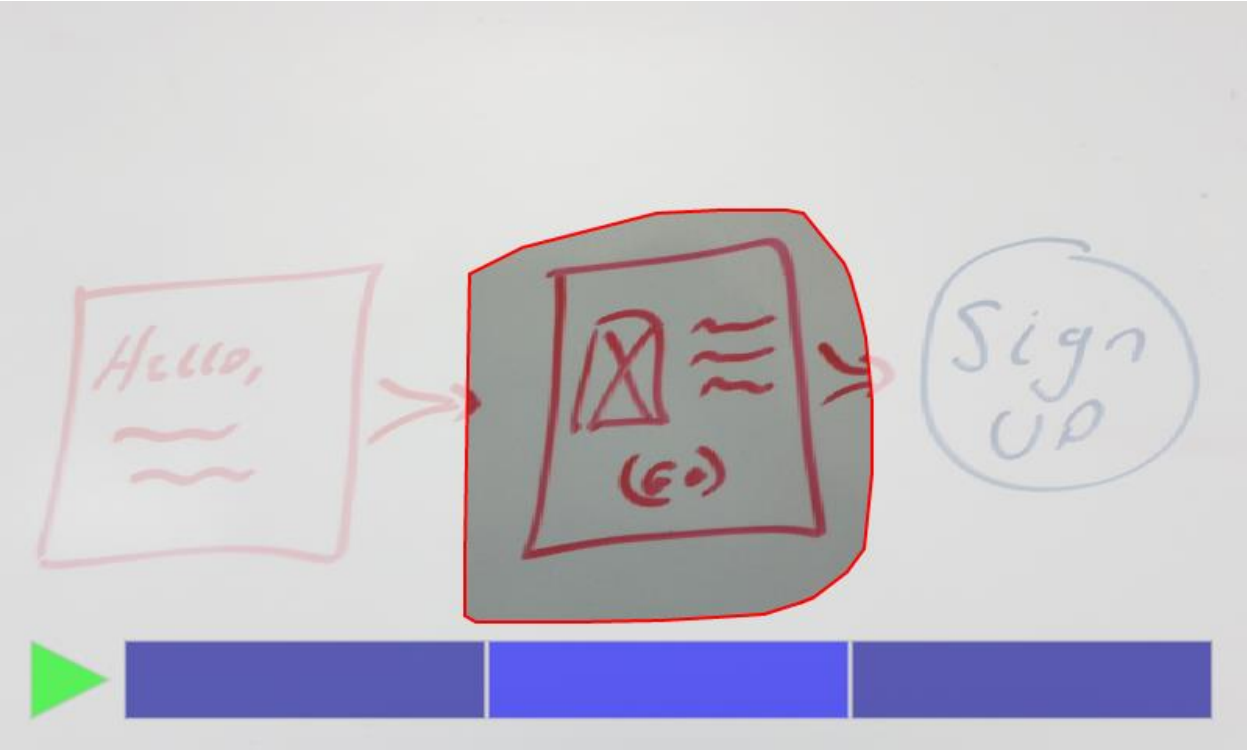

Figure 1a

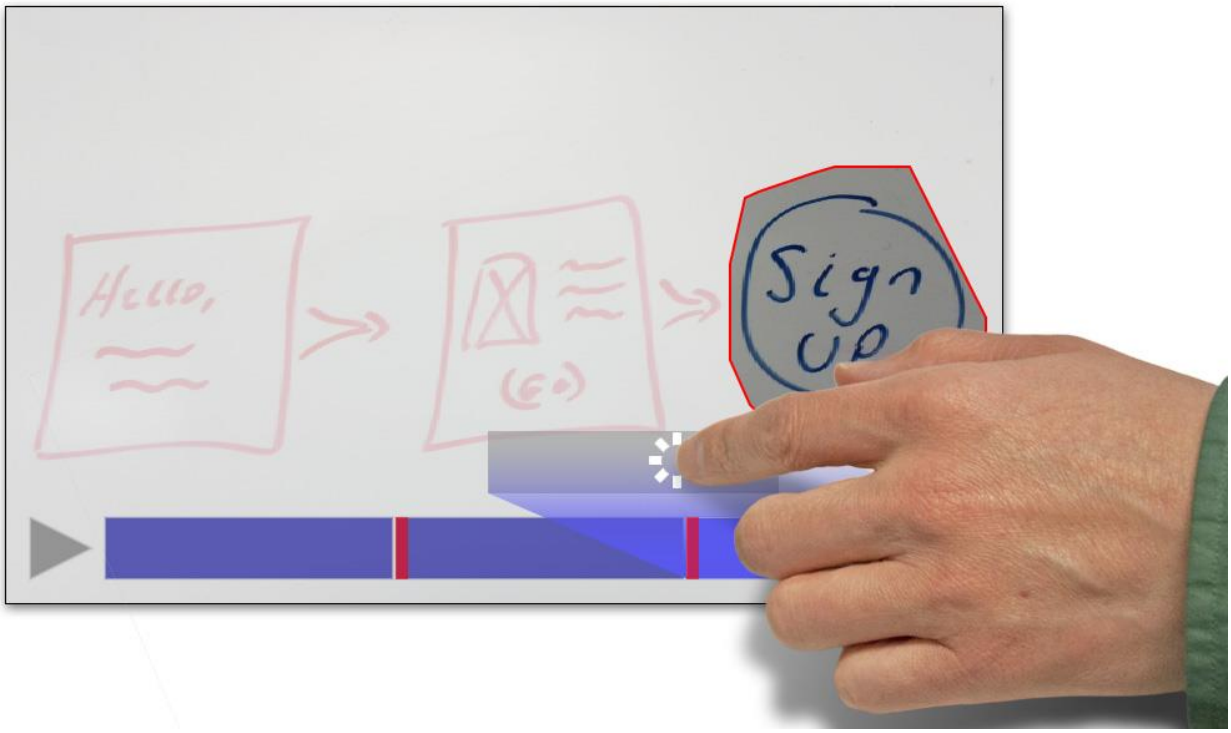

Figure 1b

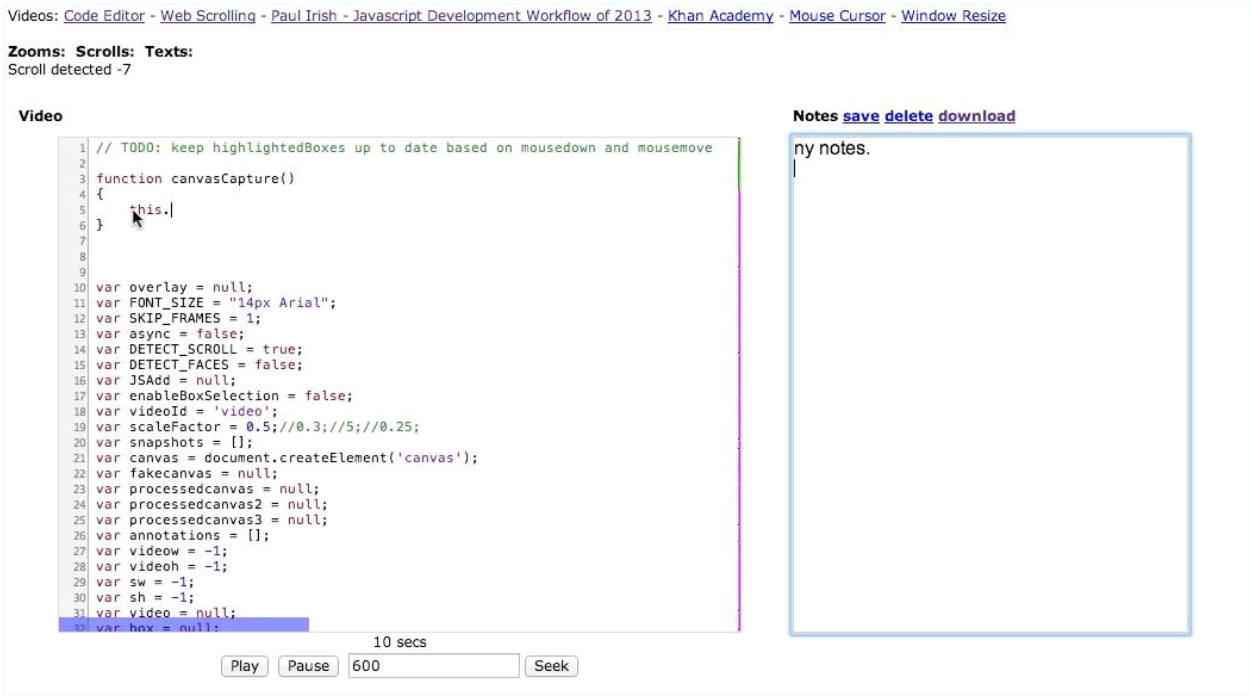

Figure 2a

Figure 2b

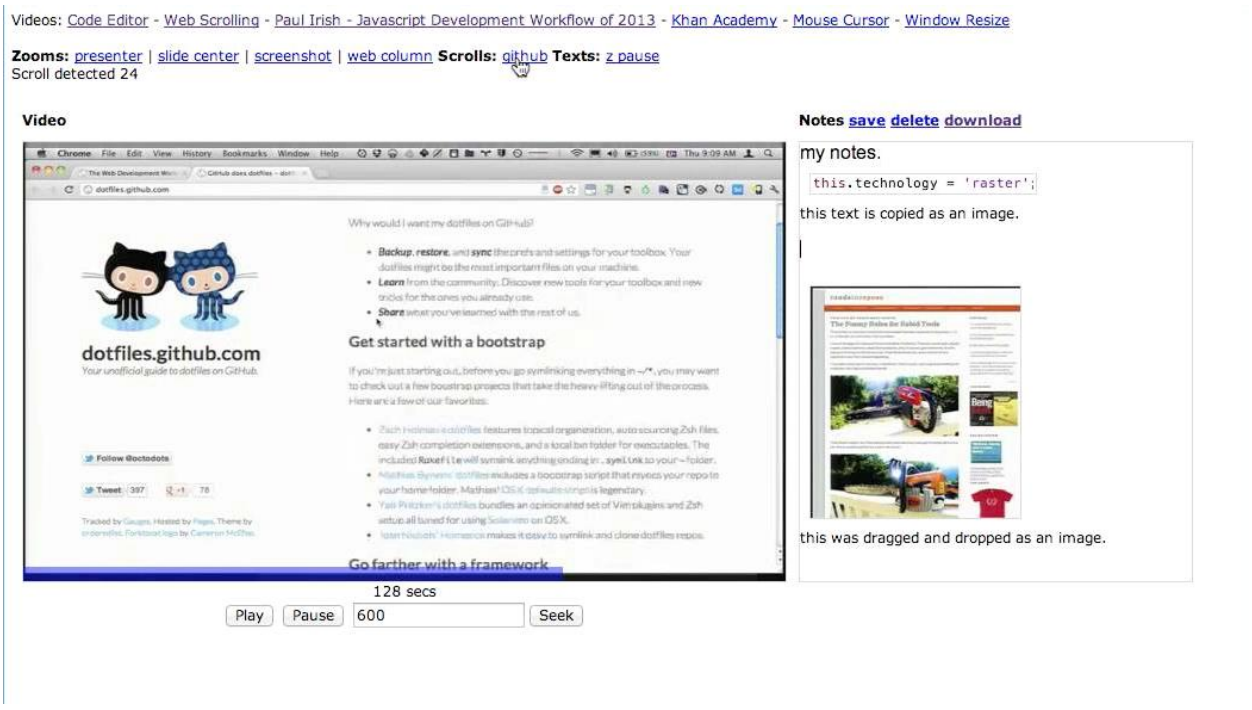

Figure 2c

Videos: Code Editor - Web Scrolling - Paul Irish - Javascript Development Workflow of 2013 - Khan Academy - Mouse Cursor - Window Resize

Zooms: presenter | slide center | screenshot | web column **Scrolls:** github **Texts:** z pause
Scroll detected 8

Video

128 secs

Play Pause 600 Seek

Notes save delete download

my notes.

```
this.technology = 'raster';
```

this text is copied as an image.

this was dragged and dropped as an image.

Figure 2d

Videos: Code Editor - Web Scrolling - Paul Irish - Javascript Development Workflow of 2013 - Khan Academy - Mouse Cursor - Window Resize

Zooms: presenter | slide center | screenshot | web column **Scrolls:** github **Texts:** z pause
scroll=0

Video

72 secs

Play Pause 600 Seek

Notes save delete download

my notes.

```
this.technology = 'raster';
```

this is copied as an image.

this is copied as an animated gif.

Figure 2e

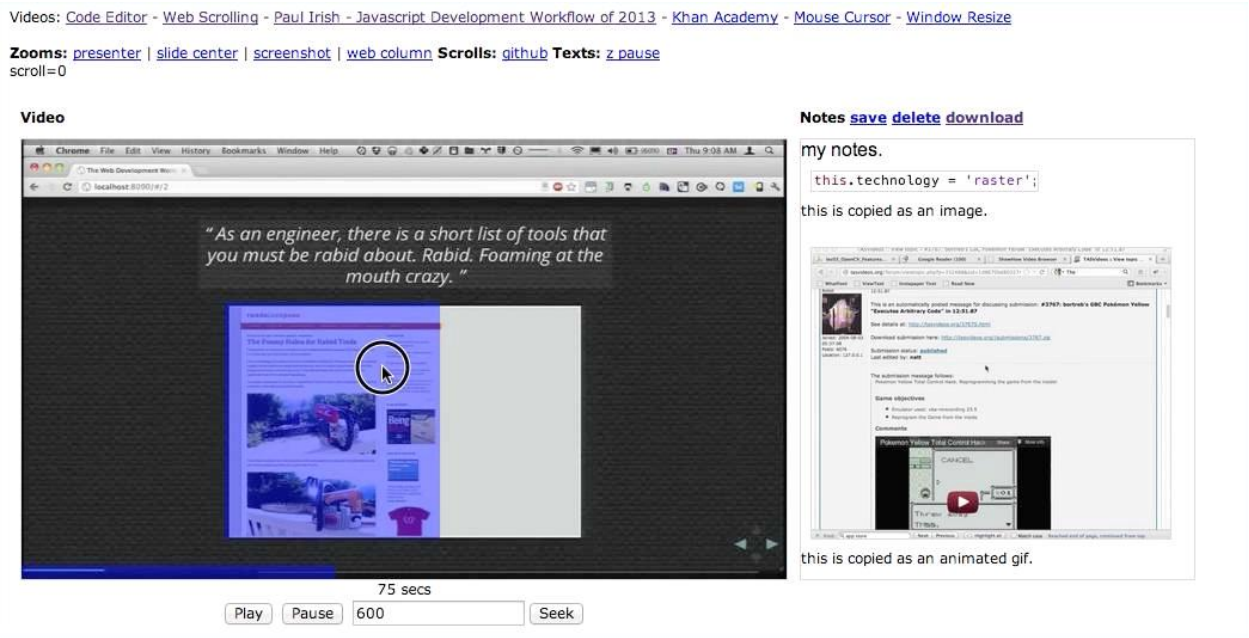
Videos: Code Editor - Web Scrolling - Paul Irish - Javascript Development Workflow of 2013 - Khan Academy - Mouse Cursor - Window Resize
Zooms: presenter | slide center | screenshot | web column Scrolls: github Texts: z pause
scroll=0
Video
"As an engineer, there is a short list of tools that you must be rabid about. Rabid. Foaming at the mouth crazy."
75 secs
Play
Pause
600
Seek
Notes save delete download
my notes.
this.technology = 'raster';
this is copied as an image.
this is copied as an animated gif.

Figure 2f