Scott Carter, Matthew Cooper, Laurent Denoue,
John Doherty, and Vikash Rugoobur,
FX Palo Alto Laboratory

Supporting Media Bricoleurs

Insights

- A media bricoleur authors new digital artifacts using whatever is at hand.
- When different types of media are integrated into a document, knowledge can be conveyed more comprehensively.
- Authoring and reuse tools for dynamic, visual media should match the power and ease of use of their static textual media analogs.

Online video is incredibly rich. A 15-minute home-improvement YouTube tutorial might include 1,500 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 percent of worldwide Internet traffic [1]. For new startups, it has become de rigueur to introduce new products with video rather than text and still graphics. This is likely because people spend more time consuming video than text. The NEA reports that Americans ages 15 to 24 watch TV two hours per day but read for leisure only seven minutes per day [2]. 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 percent more than last year [2011]... 100 hours of video are uploaded to YouTube every minute" [3]. Video is undeniably an increasingly prominent consumer communication medium.

However, it is our observation that video is not widely perceived as a full-fledged document, dismissed as a medium that, at worst, gilds over substance and, at best, simply augments text-based communications.

ILLUSTRATION BY HOROSCOPE / SHUTETRSTOCK.COM

"Idiot box" and "boob tube" are listed as synonyms for television in the *Merriam-Webster* dictionary. Even educational videos found in MOOCs have been derided as "unsophisticated chunks" [4]. But there is no overwhelming evidence that static media better convey knowledge or engender higher-quality thinking than temporal or mixed media. Indeed, humans "were never born to read"—visual storytelling predates the written word by thousands of years [5]. Furthermore, it stands to reason that marrying abstract analysis in one medium with a medium verisimilitudinous with the described content can facilitate a broader understanding of concrete concepts. 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., the scientific video site JoVE) or interactive demonstrations alongside descriptions of computational concepts (e.g., Bret Victor's Learnable Programming). Here, 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 and 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 [6]. Text documents are made up of words, which are natural compositional features endowed with objective meanings. In contrast, prevalent feature representations of visual and auditory content 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

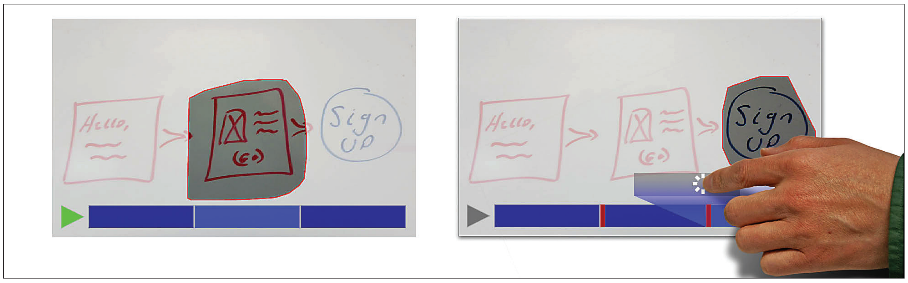


Figure 1. SketchScan overview screen, with the second of three bookmarks selected (left). The bookmark includes a region of a static image as well as an audio clip. Users can rearrange the order of clips (right). When users are satisfied with their bookmarks and annotations, they send the data to a server,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

their semantics can be relatively primitive in comparison with textual analogs.

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 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., A. Weilenmann et al. [7]). While this work is valuable, it limits media to a particular representation.

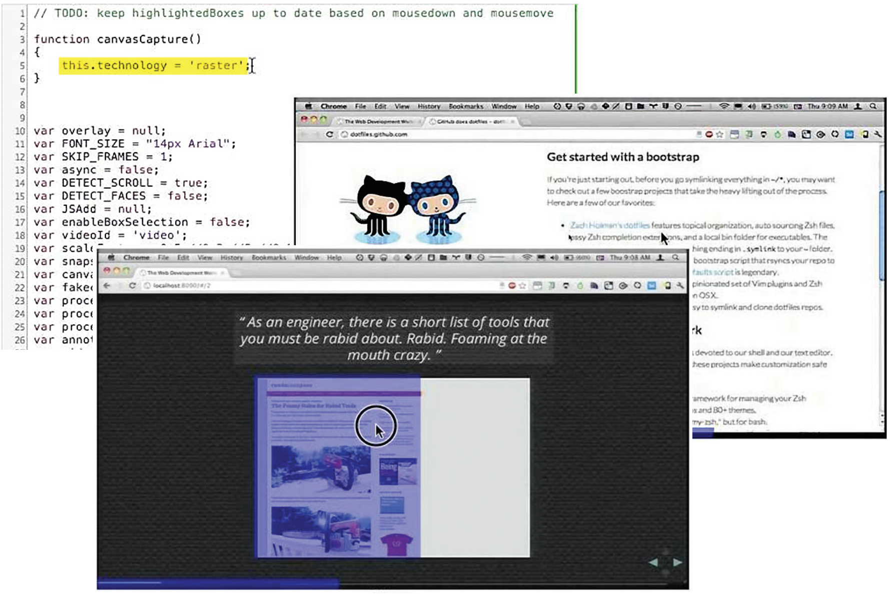
In his book *Mindstorms*, Seymour Papert suggests that "in the most fundamental sense, we, as learners, are all bricoleurs" and that we build our understanding of complicated processes by tinkering and reconfiguration. But in order to tinker you need building blocks, fully ready-to-hand components so that learners and creators can engage in what Lévi-Strauss called the bricoleur's "dialogue with…materials." 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." Multimedia documents have been supported somewhat (e.g., wikis), but these tools are not conceptually different from HTML—they still require users to mark up 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—websites 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. Text documents are designed to be consumed spatially, 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 time-based media. Several Web-based journalism sites have been exploring this approach. For example, in ESPN's long-form piece on the Iditarod, 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. 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 time point (SMIL-based authoring tools have supported similar features for many years).

We can expand the idea of responsive documents more broadly to 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 flow easily between media types so it can be tightly integrated. We are currently developing a suite of tools to support such seamless intermedia 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


```
// TODO: keep highlightedBoxes up to date based on mousedown and mousemove

function canvasCapture()
{
    this.technology = 'raster';
}



var overlay = null;
var FONT_SIZE = "14px Arial";
var SKIP_FRAMES = 1;
var async = false;
var DETECT_SCROLL = true;
var DETECT_FACES = false;
var JSAdd = null;
var enableBoxSelection = false;
var videoId = 'video';
var scal
var snap
var canv
var fake
var proc
var proc
var proc
var anno
```
Chrome File Edit View History Bookmarks Window Help
Thu 9:09 AM
The Web Development Wo
Github does dotfiles – dot
dotfiles.github.com

Get started with a bootstrap

If you're just starting out, before you go symlinking everything in ~/*, you may want to check out a few boostrap projects that take the heavy lifting out of the process. Here are a few of our favorites:

- Zach Holman's dotfiles features topical organization, auto sourcing Zsh files, easy Zsh completion exte ons, and a local bin folder for executables. The hing ending in .symlink to your ~ folder. bootstrap script that rsyncs your repo to faults script is legendary. pinionated set of Vim plugins and Zsh n OSX. sy to symlink and clone dotfiles repos.

rk

devoted to our shell and our text editor. hese projects make customization safe

amework for managing your Zsh s and 80+ themes. my-zsh," but for bash.

"As an engineer, there is a short list of tools that you must be rabid about. Rabid. Foaming at the mouth crazy."

Figure 2. Directly interacting with video content with Cemint. Users can highlight text (top), manipulate the mouse wheel to scroll (middle), and select regions of importance (bottom) to crop the video or to copy content to their personal notepad.

used to capture, clean, animate, and share sketches, is a demonstration of the former [8]. With this app, users define regions of a sketch, optionally add audio annotations to each region, and ultimately generate 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 [9]. But there are many other ways to leverage video content in user interfaces. As part of Cemint, we are building tools that allow users to extract any keyframe from a video, or automatically detected subregions 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, using the mouse wheel to scroll up and down, or double-clicking to identify rectangular areas of importance (Figure 2). Users then use familiar copy-and-paste or drag-and-drop techniques to extract content to multimedia documents [10].

One effect of supporting the flexible repurposing of content across media is freeing users to compose thoughts in the domain of their choice 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, for example, 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 [11]. Furthermore, content analysis can support users' compositions. Analysis can leverage user input solicited via familiar interactions with both the original content and exposed intermediate results of real-time analysis. This user-driven approach to content analysis can avert many difficulties that plague the predominant automatic end-to-end analysis paradigm.

We are just beginning our work in this area—we are far from providing full-fledged multimedia document support. And there are many other ways to apply text 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 from video conferences or lectures, can enable better note-taking, review, and content reuse. Tools or techniques that make it easy for spatial navigation to trigger side effects that enrich the reading experience without detracting from the comprehension of main concepts represent 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. As David Weinberger writes, "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" [12].

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.

Endnotes

1. http://technews.tmcnet.com/iptv/topics/iptv/articles/136034-video-dominates-global-traffic.htm
2. NEA. To Read or Not To Read: A Question of National Consequence. 2007.
3. http://www.youtube.com/yt/press/statistics.html
4. Vardi, M.Y. Will MOOCs destroy academia? *Comm. of the ACM 55*, 11 (2012), 5.
5. Wolf, M. *Proust and the Squid: The Story and Science of the Reading Brain*. Harper Perennial, 2008.
6. Hauptmann, A., Yan, R., Lin, W-H., Christel, M., and Wactlar, H. Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news. *IEEE Transactions on Multimedia 9*, 5 (2007), 958–966.
7. Weilenmann, A., Säljö, R., and Engström, A. Mobile video literacy: Negotiating the use of a new visual technology. *Personal and Ubiquitous Computing 18*, 3 (2014), 737–752; http://dx.doi.org/10.1007/s00779-013-0703-x
8. http://sketchscan.fxpal.com/
9. Uchihashi, S., Foote, J., Girgensohn, A., and Boreczky, J. Video Manga: Generating semantically meaningful video summaries. *Proc. of the ACM International Conference on Multimedia*. 1999, 383–392.
10. Denoue, L., Carter, S., and Cooper, M. 2013. Content-based copy and paste from video documents. *Proc. of the ACM Symposium on Document Engineering*. 2013, 215–218.
11. Huh, J. Why Microsoft Word does not work for novice writers. *Interactions 20*, 2 (2013), 58–61.
12. Weinberger, D. *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.

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

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.
→ cooper@fxpal.com

Laurent Denoue is a researcher at FX Palo Alto Laboratory interested in user interaction design and document and video processing. He worked on XLibris, an annotation system; ProjectorBox, an appliance for capturing meetings; and TalkMiner, a service that detects slides in online lectures. His recent interest is client-based video processing to manipulate video documents in real time.
→ denoue@fxpal.com

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.
→ doherty@fxpal.com

Vikash Rugoobur is a visiting researcher at FX Palo Alto Laboratory. His interests include quantified self, wearable devices, and user interaction.
→ vik@fxpal.com

DOI: 10.1145/2617379 COPYRIGHT HELD BY AUTHORS. PUBLICATION RIGHTS LICENSED TO ACM. $15.00