ShowHow: Supporting Expository Video Capture and Access

要　旨

ビデオ・コンテンツ制作者は膨大な労力を費やして作品を制作するが、それを受け取る側は受動的に視聴することが一般的である。しかしながら、ビデオを用いて学習を行う場合は、コンテンツを視聴するだけではなく、コンテンツに関与し、相互作用しながら、違う目的で活用できるように、合わせ込みを行うことがユーザーに求められる。さらに、コンテンツ製作者がビデオ撮影者である必要がない分野で、ビデオを用いた学習を促進するためには、データを取り込むツール群がインタラクティブ・コンテンツの作成を助けることが重要である。本論文ではこのゴールに向けて行った初期の実験について説明する。デザインのための形成的なフィールド・スタディーと、文献レビューから、一連の広範なビデオ作成と、相互作用のスタイルを取り入れることができるシステムのデザインに繋がった。

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. A literature review coupled with formative field studies led to a system design that can incorporate a broad set of video-creation and interaction styles.

Author

Scott Carter
Matthew Cooper
John Adcock

FX Palo Alto Laboratory, Inc.

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. Amid the growth of Internet sharing and pervasive mobile devices, the mass creation of online expository videos, including how-to, tutorial, and lecture videos, is an emerging trend. In this work, we explore augmenting 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 [14], graphics that show how to accomplish a task improve understanding beyond textual descriptions [9].

For some tasks, video has been shown to be particularly helpful beyond static graphics [8,16]. 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 involve 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 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 understanding past work investigating the role of media in learning and knowledge transfer and conducting first-hand, participatory observations of the use of off-the-shelf capture and access tools.

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 primary task

performance but poor learning and transfer to other tasks. 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 [15] and that learning can improve when complemented with video-based examples in particular [13]. 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. For this, tools should support users actively navigating [17] as well as annotating and editing to develop their own interpretations of video content [21,3]. 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 [22].

Overall, 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.

We found similar issues in our formative field work. Our observations (reported more extensively here [4]) showed that how-to video authors require straightforward, unobtrusive capture. We also found that how-to creators and users alike want to be able to add marks and multimedia annotations to video and to share their videos and annotations with colleagues.

In summary, Torrey et al. found that how-to “sharing occurs within and across a collection of communication tools without any centralized control” [18] and that people tend to find information by browsing as much as by more directed search [19]. 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 consistent view. Content creators and learners alike can use a single web-based application to browse, skim, and navigate 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:

Capture Users can upload arbitrary videos to our server via a simple drag-and-drop interface, or they can specify a URL of a YouTube video to annotate. We have also built tools to convert different types of expository content into annotateable videos (see 3.1).

View and annotate Inspired by other tools that use bookmarks to index interface actions [2] and other events[*1], we built an HTML5-based web client to support creation, editing, sharing, and viewing of bookmarked, annotated videos (see Figure 1). This client is designed for desktops and tablets and supports bookmarking and annotation. Users can also filter bookmarks by content-creator or with a live search of the bookmark's title and text annotations.

[*1] http://teachscape.com

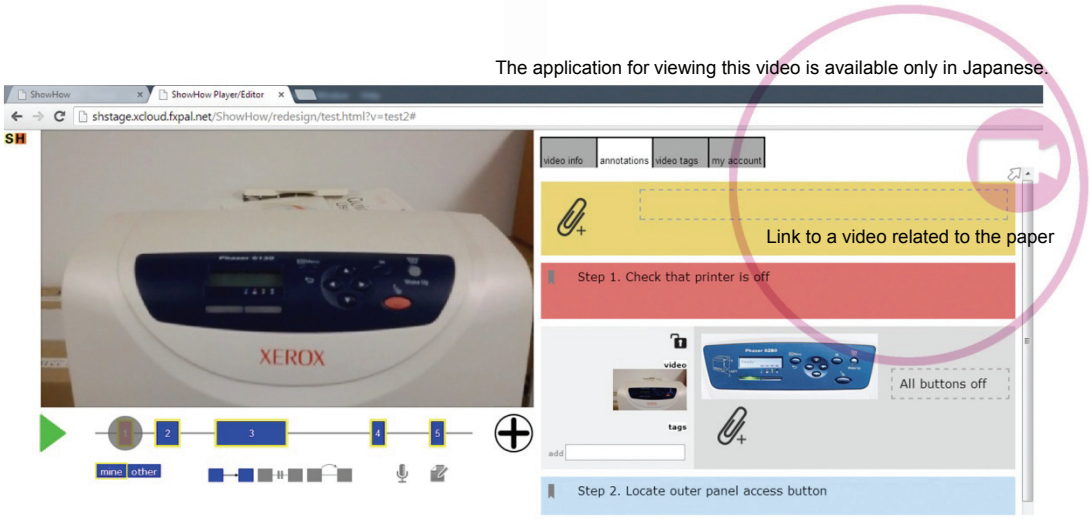


Figure 1. The main ShowHow web viewer and editor. (Left) Numbered rectangles along the timeline are bookmarks. Users can toggle the player to pause after each bookmark or skip material between bookmarks; re-record the audio track; or edit bookmark durations. (Right) The annotation view allows users to label bookmarks as well as to add multimedia annotations. Here, the user has added a labeled photo annotation just below the first bookmark.

Mobile viewer We 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 bookmarks and annotations are searchable in a separate interface. Search results can link to an entire video or a specific bookmark within a video.

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

3.1 Bookmarking informal videos

For more informal expository content, such as guides or how-to videos, bookmarks are typically either explicit steps or important points within the video. Bookmarks in ShowHow can be used flexibly. The bookmarks in Figure 1 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, SketchScan[*2] 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. We built the SketchScan system before ShowHow, but the content it yields lends itself easily to bookmarking. Without changing SketchScan 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 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 created an ingest tool to ShowHow that converts each step, which may

*2 http://sketchscan.fxpal.com

be a video clip or a photo and may include optional text, into a single video with text-based annotations. This conversion process allows us to view yet another type of how-to content in a consistent interface.

3.2 Bookmarking produced videos

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

In this domain, we have thus far focused on lecture content available through Coursera[*3], which recently enrolled its one millionth student[*4], and specifically the Machine Learning course[*5]. The course's video content blends shots of the instructor speaking and shots of a slide stream with audio commentary. The videos 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 reflect the presenter's topical structuring of the content. We 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.

The second component is a simple frame difference detector adapted from previous work on lecture video analysis [1]. We compare temporally adjacent frames pixel-wise. We filter out regions of change with low spatial support, and sum the remaining number of pixels above a threshold. When the number of changed pixels exceeds 45% of the frame, 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 avoids detecting multiple incomplete versions of a single annotation.

Given a source video, we first apply the SVM classifier to detect frames that show 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 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 2 shows an example in which the

[*3] http://coursera.com
[*4] http://goo.gl/tmWMF6
[*5] https://www.coursera.org/course/ml

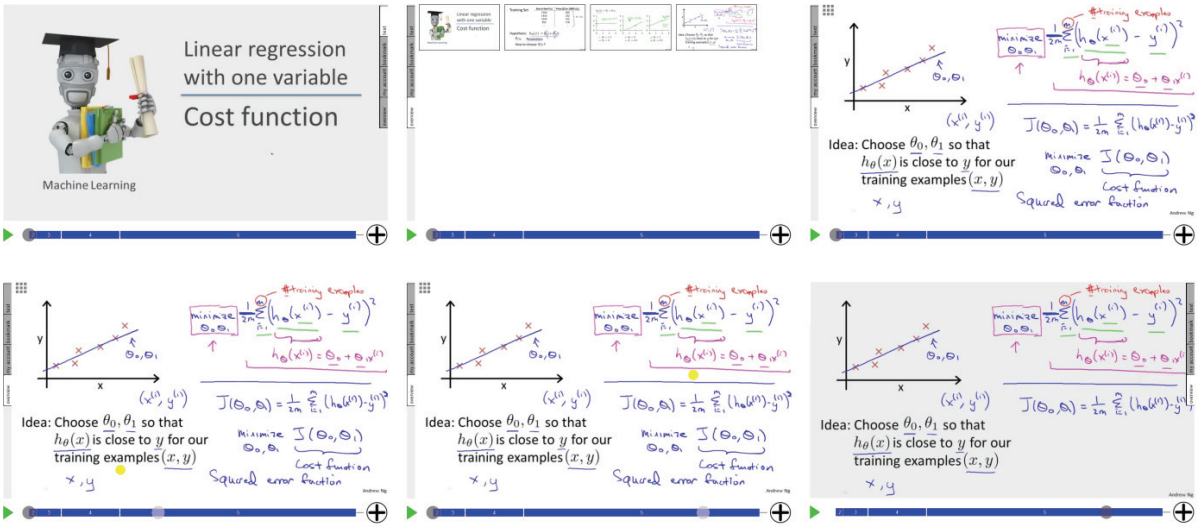


Figure 2. The ShowHow UI viewing a Coursera video on a tablet using overview navigation. Bookmarks on the timeline correspond to slides, each of which have a separate overview image. The user expands the overview image tab (top left and center) and selected an overview image (top right). As the user drags across the overview image (the cursor location is shown in yellow) a secondary indicator on the timeline (light gray circle) shows the creation time of the selected content (bottom left and center). When the user lifts up from the timeline the system navigates the video to the selected time, hides the overview image, and begins playing (bottom right). Swiping the screen navigates between overview images.

slide-based segments appear as bookmarks for one of the Coursera videos. 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.

3.3 Spatial navigation

ShowHow can exploit the detected slide segmentation to generate navigable overviews with which users can efficiently “skim” the video.

After detecting slide changes events and the subordinate annotation boundary boxes, the system generates interactive overview images so that users can navigate the video spatially. The overview image displays the complete set of annotations.

The bounding boxes of slide annotations can enhance navigation of the overview image. The system uses hierarchical clustering of the set of detected bounding boxes to experiment with the number (granularity) of time points at which the user can jump into the video. Grouping by time is the most natural approach. However, incorporating spatial information into the clustering is a natural extension when the content is added in a consistent manner such as “left to right” or “top to bottom”.

We built an interface in which all distinct slides appear and are shown with all added presenter annotations, providing hierarchical non-linear access. Users can first indicate a slide segment of interest. By selecting an annotation shown on the overview thumbnail for that slide segment, users can navigate to the sub-segment in which the annotation is added.

An example of this class of content appears in Figure 3. Figure 3 (left, center) shows the first and last frames from an automatically detected slide segment. The addition of both slide text and presenter annotations is evident through the course of the time the slide appears in the video. Figure 3 (right) shows the results of annotation detection and clustering. Each cluster of annotations (indicated by the color of the overlaid bounding boxes) partitions the slide segment in time into sub-segments. Users can directly access the sub-segments by selecting the annotations allowing them a second level of non-linear access to the video.

Once constructed, overview images are integrated into our video player (see Figure 2). In the UI the timeline is divided into segments corresponding to each boundary event (i.e., the

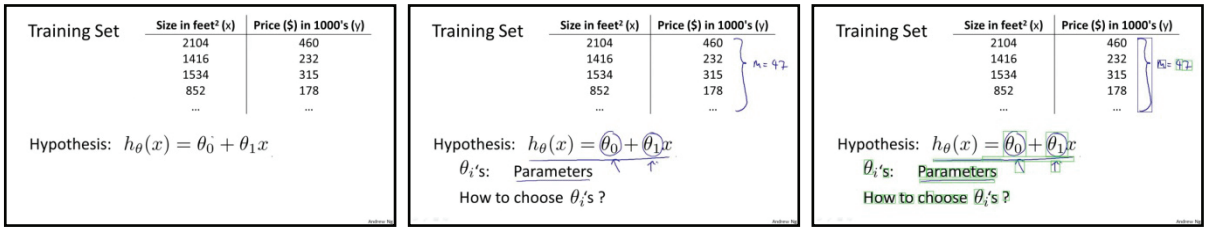


Figure 3. Analysis flow for a Coursera video. (Left) shows the first frame in a slide segment. (Center) shows the last frame from the same slide segment 37 seconds later. Notice the addiction of both slide text and written annotations. (Right) shows detected annotations using bounding boxes. The color of the boxes indicates cluster membership. The clusters partition the slide segment into subshots allowing for fine grained access into a slide segment using spatial navigation (Figure 3).

timespan of each overview image). When interacting with a navigable overview with sub-segments, pressing and holding a region of the overview image will generate feedback on the timeline indicating to what time the video would seek if the user were to click on that region. The feedback will update automatically as the user moves their finger (or a mouse cursor) around the overview image. When they release, the system navigates to the indicated position after a delay.

This system can be applied to video exhibiting content change events other than slide transitions. For example, content in Khan Academy[*6] videos transitions via scrolls – our system can detect these scrolls and build navigable overviews similar to the slide-based content described above (and similar to [12]).

4. Conclusion and future work

Past systems for disseminating knowledge focused on textual descriptions of problems and solutions [20]. In contrast, our goal 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. Literature reviews and case studies we conducted led us to the conclusion that the best tools are those that flexibly incorporate a variety of tools that support different styles of capture and access. This directly inspired 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 suggested that first-person video instructions can improve performance on assembly [10] and learning [11] 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.

[*6] http://www.khanacademy.org

5. TRADEMARKS

- YouTube is a registered trademark of Google Inc.
- Coursera is a registered trademark of Coursera, Inc.
- Khan Academy is a trademark of Khan Academy.
- All brand names and product names are trademarks or registered trademarks of their respective companies.

6. References

[1] Adock, J., Cooper, M., Denoue, L., Pirsiavash, H. and Rowe, L.A. Talkminer: A lecture webcast search engine. ACM MM. pp. 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. pp. 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). pp. 59-75. (2013).

[4] Carter, S. Adcock, J., Cooper, M., and Branham, S. Tools to support expository video capture and access. Education and Information Technologies Journal. DOI: http://dx.doi.org/10.1007/s10639-013-9276-6. (2013).

[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. pp. 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). pp.749-770. (2011).

[8] Grossman, T. and Fitzmaurice, G. ToolClips: An investigation of contextual video assistance for functionality understanding. ACM CHI. pp. 1515-1524. (2010).

[9] 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. pp. 82-89. (1995).

[10] Kraut, R. E., Fussell, S. R. and Siegel, J. Visual information as a conversational resource in collaborative physical tasks. Human-Computer Interaction. 18(1). pp. 13-49. (2003).

[11] Lindgren, R. Generating a learning stance through perspective-taking in a virtual environment. Computers in Human Behavior. 28(4). pp. 1130-1139. (2012).

[12] Monserrat, T-J.K.P., Zhao, S., McGee, K. and Pandey, A.V. NoteVideo: Facilitating navigation of blackboard-style lecture videos. ACM CHI. pp. 1139-1148. (2013).

[13] 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). pp. 449-465. (2008).

[14] 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). pp. 687-701. (1991).

[15] Pirolli, P. Effects of examples and their explanations in a lesson on recursion: A production system analysis. Cognition and

Instruction. 8(3). pp. 207-259. (1991).

[16] 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. pp. 135-144. (2011).

[17] Schwan, S. and Riempp, R. The cognitive benefits of interactive videos: Learning to tie nautical knots. Learning and Instruction. 14(3). pp. 293-305. (2004).

[18] Torrey, C., McDonald, D., Schilit, W. and Bly, S. HowTo pages: Informal systems of expertise sharing. ECSCW. pp. 391-410. (2007).

[19] Torrey, C., Churchill, E. F. and McDonald, D. W. Learning how: The search for craft knowledge on the internet. ACM CHI. pp. 1371-1380. (2009).

[20] 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. pp. 257-284. (2011).

[21] 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). pp. 403-440. (2010).

[22] 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). pp. 15-27. (2006).

Author's Introductions

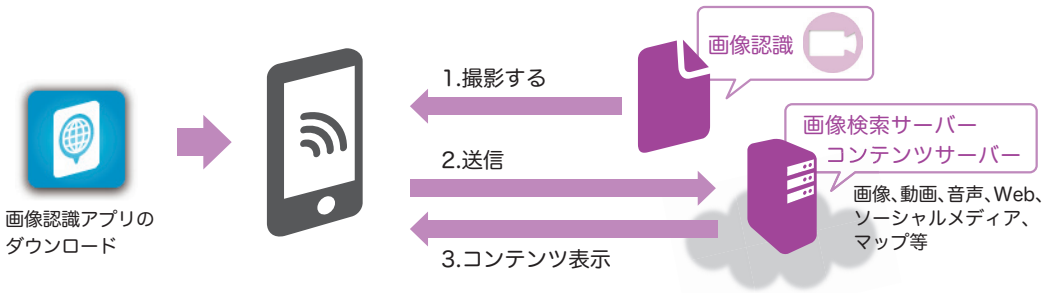
Scott Carter
FX Palo Alto Laboratory
Area of specialty: Computer Science (Ph.D.), Human Computer Interaction

Matthew Cooper
FX Palo Alto Laboratory
Area of specialty: Electrical Engineering (Ph.D.), Multimedia and Machine Learning

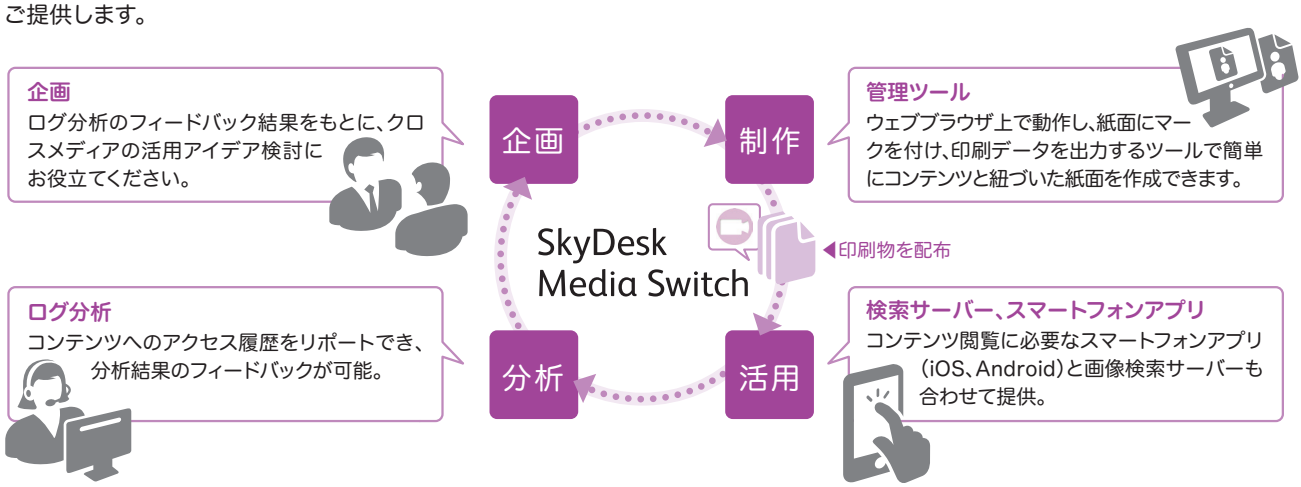
John Adcock
FX Palo Alto Laboratory
Area of specialty: Electrical Engineering (Ph.D.), Multimedia Systems

「富士ゼロックス テクニカルレポート」は SkyDesk Media Switch に対応しています。

SkyDesk Media Switchは画像認識技術を使ったクロスメディアサービスです。スマートフォン／タブレットで紙から簡単に動画などのマルチメディアコンテンツを再生できます。（日本語のみ対応しています。Available only in Japanese）



企画→制作→活用→分析、そしてまた企画というサイクルを効率的に実践し、改善していくために必要なツールをオールインワンでご提供します。



詳しくはSkyDesk Media Switchのサイトをご覧ください！ Media Switch 検索

専用アプリで関連情報にアクセス！

「富士ゼロックス テクニカルレポート」は、App Store・Google Play から SkyDesk Media Switch のアプリ（ダウンロード無料）をインストールし、アプリを起動したスマートフォンで紙面の特定画像を撮影すると、各関連情報にアクセスいただけます。

＊対象 OS（iOS）: iOS 6.0、6.1、7.0、Android™ 2.3.x、Android™ 4.0.x、Android™ 4.2.x
＊アクセスできる動画のリンク先は、予告なく閉鎖される場合がありますので、予めご了承ください。

◆アプリのインストールの手順

App Store/Google Play にアクセス　「Media Switch」で検索　インストールを実行

◆アプリのご利用手順

アプリを起動
検索窓に「富士ゼロックス」を入力
結果から「富士ゼロックス テクニカルレポート」を選択
カメラアイコンをタップして撮影モードに移動
画像全体がガイドの中に入るようにして撮影
送信するとコンテンツへアクセス