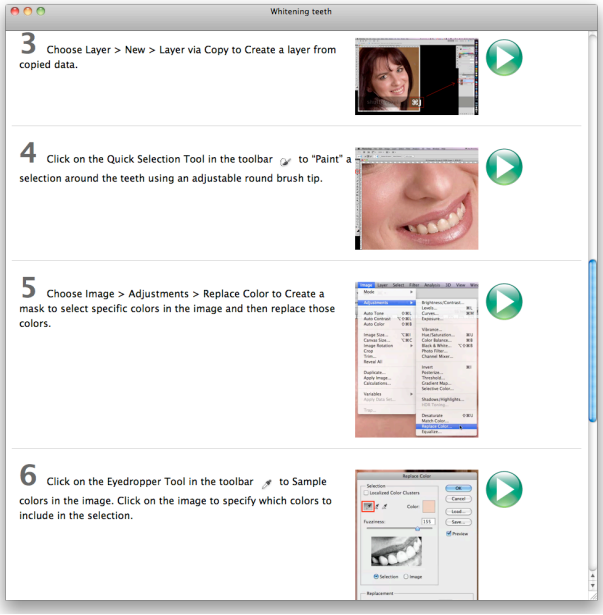


figure 1. Partial view of the MixT interface, which shows step-by-step instructions, each with a screenshot and a video clip.

MixT: Automatic Generation of Step-by-Step Mixed Media Tutorials

Pei-Yu (Peggy) Chi
University of California, Berkeley
533 Soda Hall
Berkeley, CA 94720 USA
peggychi@cs.berkeley.edu

Björn Hartmann
University of California, Berkeley
533 Soda Hall
Berkeley, CA 94720 USA
bjoern@cs.berkeley.edu

Sally Ahn
University of California, Berkeley
533 Soda Hall
Berkeley, CA 94720 USA
sallyahn@berkeley.edu

Mira Dontcheva
Adobe Systems
601 Townsend Street
San Francisco, CA 94103 USA
mirad@adobe.com

Amanda Ren
University of California, Berkeley
533 Soda Hall
Berkeley, CA 94720 USA
aren@berkeley.edu

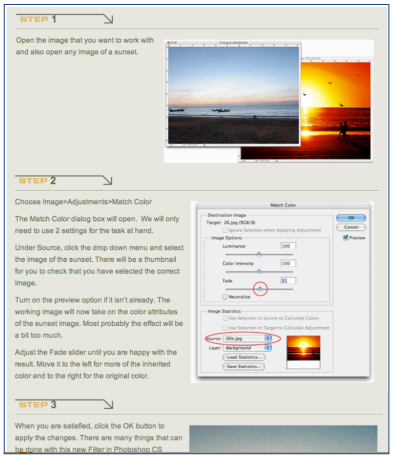
Wilmot Li
Adobe Systems
601 Townsend Street
San Francisco, CA 94103 USA
wilmotli@adobe.com

CHI'12, May 5–10, 2012, Austin, Texas, USA.
ACM 978-1-4503-1016-1/12/05.

Abstract
As software interfaces become more complicated, users rely on tutorials to learn, creating an increasing demand for effective tutorials. Existing tutorials, however, are limited in their presentation: Static step-by-step tutorials are easy to scan but hard to create and don't always give all of the necessary information for how to accomplish a step. In contrast, video tutorials provide very detailed information and are easy to create, but they are hard to scan as the video-player timeline does not give an overview of the entire task. We present MixT, which automatically generates mixed media tutorials that combine the strengths of these tutorial types. MixT tutorials include step-by-step text descriptions and images that are easy to scan and short videos for each step that provide additional context and detail as needed. We ground our design in a formative study that shows that mixed-media tutorials outperform both static and video tutorials.

Author Keywords
Workflow capturing; tutorials; instructions; videos; screencast

ACM Classification Keywords
H.5.m [Information Interfaces and Presentation (e.g., HCI)]: Miscellaneous.



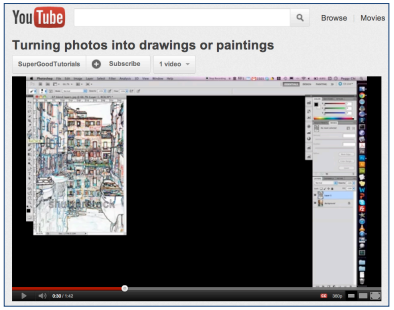


figure 2. Today, most tutorials are presented as either static content, e.g., on web pages (top), or as continuous videos, e.g., on YouTube (bottom).

Introduction

Tutorials are learning materials that offer step-by-step instructions on how to perform a task. Currently, tutorials for learning software are presented in two dominant forms:

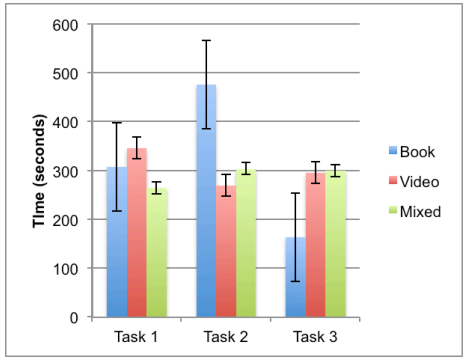
- **Static tutorials**, a mix of text and images found in books or on the web (figure 2 top), offer quick access to information because their step-by-step nature facilitates scanning content that is presented all at once. However, static images only provide visual information of certain states, which is sometimes not enough to teach complex techniques involving many interactions or long, continuous actions, such as brushing a region, drawing a selection path, adjusting multiple control points, or rotating a 3D object.
- **Full-length Videos**, on the other hand, reveal the relationship between users' actions and system response [7] (figure 2 bottom), which helps to demonstrate an entire process. The disadvantage, however, is that linear video for users to find specific steps or actions within a larger tutorial.

Both static and video tutorials have their own advantages for different learning goals: Studies have shown that animated demonstrations yield faster and more accurate performance during learning sessions in which users "mimic" the actions to acquire new skills, and static text instructions yield better performance in later sessions that require users to recall the learned skills [6]. We draw inspiration from such findings and hypothesize that a combination of video and static instructions can improve both the acquisition and retention of new skills through tutorials. We target image editing software such as Adobe Photoshop in particular because it is widely used and has a large collection of tutorials accessible in bookstores, video platforms (e.g. YouTube), and on the web (e.g. Photoshop Gurus Forum). In this way, users may effectively learn complicated actions (e.g. applying brush strokes) from tutorial video clips, and quickly access simple actions (e.g. copying a layer) from static text and images.

Related Work

Researchers have explored various ways to construct learning materials by capturing expert demonstrations. Grabler *et al.* presents a system that automatically creates tutorials from recorded demonstrations of application usage [3] and later learns the parameters for image editing operations and generates macros that takes into account the context of the current image [1]. Chronicle captures video and history of graphical documents to create an interactive learning tool that offers video playback and visualization of user actions [5]. There are also research projects aiming to enrich in-application tutorials with multimedia. ToolClips enhances traditional tooltips by including contextual text and video information of application commands for users to see the relevant resources while navigating the interface [4]. Pause-and-Play tracks user actions and supports online video tutorial playback based on the events users are performing, to avoid users switching back and forth between user context and online tutorials [8]. These projects have looked at how to combine various kinds of mixed-media tutorials but have not investigated how different types of media support particular actions. This motivates us to investigate whether correlations exist between types of media and types of commands, and whether they can be used to inform the automatic generation of our mixed-media tutorials.

(a) Mean Task Completion Times



(b) Total # of Errors and Missed Steps

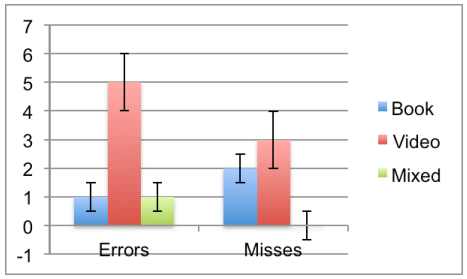


figure 3. Quantitative results, including completion time, errors and misses, of the preliminary study.

Initial Study

Based on previous studies that showed how computational step-by-step static tutorials performed better than existing static presentations [3] and advantages exist for both static and animated-demonstration methods [6], we designed a pilot study to investigate whether mixed-media tutorials help users follow a task, and if so, whether videos benefit users for certain commands. Our initial study aims to evaluate the following two hypotheses:

H1 Image manipulation tutorials that mix static images and video clips are more effective than all-static or all-video tutorials.

H2 There are certain types of commands and tasks where users benefit from seeing video clips instead of static text and images.

We recruited 4 participants (1 male and 3 females, aged 21-24) from a campus student design group, all undergraduate students. Since our tutorials focused on achieving specific tasks rather than introducing new users to the software, we wanted participants who were familiar with the basic functions and navigation of Photoshop, but not expert enough to perform the specific effects. Therefore, our participants had between 4 to 9 years of experience using Photoshop.

Our experiment was based on within-subject design. We chose 3 different image manipulation tasks that represent similar levels of difficulty and complexity consisting of 10-15 steps: 1) whitening teeth, 2) transforming a photograph to a watercolor image, and 3) blending two images to add depth of field. Each participant was asked to perform these 3 tasks by following the tutorials in a random sequence of assigned type: book, video (on YouTube), or mixed (a web prototype of the interface shown in figure 1). To ensure the information was equally presented, we followed the book tutorials used for comparison to record and narrate our video tutorials, and manually generated our mixed tutorials. We modified some textual instructions of the book tutorial to more closely match our video and mixed tutorials. Each session consisted of 3 tasks and was 30 minutes long. The study was conducted in a lab environment, using laptop computers running Mac OS X and Adobe Photoshop CS5.1, a web browser, and a provided mouse.

To answer **H1**, we gathered quantitative data: the users' completion time and number of errors. We counted an *error* as any incorrect or extraneous commands that users performed during their task. We counted a *miss* as any steps in the tutorial that the user accidentally skipped. It's worth noting that skipping a crucial step might lead the user to perform several redundant steps (counted as one error), which he must later undo in order to complete the task. We collected this data by capturing screencasts of all users' actions and analyzing the video afterwards. We also collected qualitative data by observing how users followed the presented information and requesting feedback via Likert-scale questionnaire. To answer **H2**, we recorded the number of video clicks in the mixed tutorial to understand which commands and tasks compel users to play the video clip.

Results

User performance of image editing tasks

Figure 3a shows the average time to complete each task, by tutorial types, with standard error. There was no significant difference in completion time between the

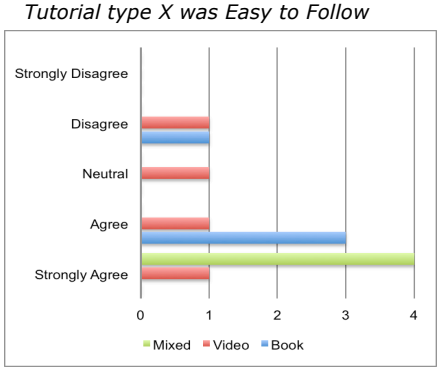


figure 4. Users' response to tutorial types

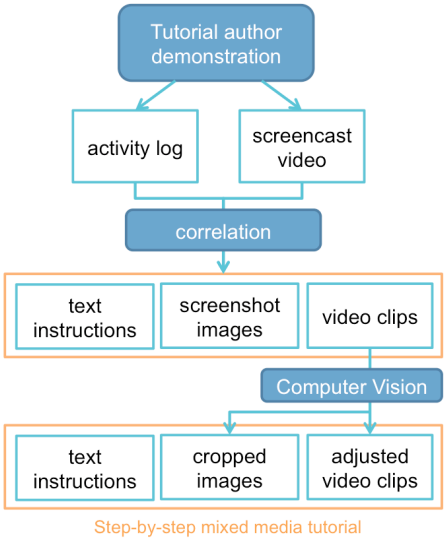


figure 5. MixT system pipeline

tutorial types (p-value>0.8 in t-test). However, mixed tutorials reduced the total number of errors compared to videos, and reduced missed steps for all tasks, as shown in figure 3b. It is interesting to note the irregular spike in completion time for the book tutorial type for Task 2, which includes a step involving the brush tool. Whereas the video and mixed tutorial demonstrated that this step should only involve quick and rough strokes in the center, one participant misunderstood the book's description and devoted nearly 5 minutes to carefully brushing in the different buildings in the image for this task. This episode exemplifies our belief that static images and text are not always sufficient for communicating tasks that involve continuous actions.

We then wanted to understand which steps and their corresponding commands prompted users to view the information in the video format (**H2**). We found three types of actions participants often viewed as video clips, which are the following:

1) Finding target buttons or manipulating UI elements, such as locating a tool on a side panel (brush tool), navigating the main menu panels (file-open, auto-blend), and setting a parameter slider (brush size, layer opacity, vibrance).
2) Performing freehand, continuous actions, such as applying brush strokes and selecting a region.
3) Examining the command effect, such as hiding/showing an image layer.

These findings were in accordance with Palmiter and Elkerton's studies on how users learn by "mimicking actions" from animated demonstration [6], and supported our belief that video clips particularly benefit users in certain types of commands. We speculate that videos assist with types 1 and 2 by explicitly demonstrating the necessary cursor movements for the required tasks rather than requiring the user to infer the needed movements from static images.

User preference of tutorial types

Results of our questionnaire show that while participants had varying opinions on the book and full-length video tutorials, all users strongly agreed that the mixed tutorial was easy to follow (Figure 4). For book tutorials, participants had difficulty finding the tools that the tutorial referenced and remarked that there were not enough visuals. For full-length videos, participants disliked having to pause the video to complete each step. For the mixed tutorial, half the participants found that video was most useful. One participant acknowledged that because the mixed tutorial allowed him to "break down the process into simple steps," he was able to easily find the point where he made a mistake. Another user explained that videos would be most helpful if the tasks were more advanced and in-depth. We believe that one advantage of a mixed tutorial over the other tutorials is that it uses a single screen to provide the user with a choice over the different information types at every step, thus giving the user easy control over how much time and attention (e.g. skimming text or playing video) to devote to each step.

Introducing MixT: Video-Mixed Tutorial

The results of this pilot led to the design of our system, *MixT*, which captures a workflow while an author demonstrates a task and automatically generates a mixed media tutorial designed to assist users in navigating instructions. Figure 1 shows our mixed tutorial interface, where a user can see a textual

(a)

(b)

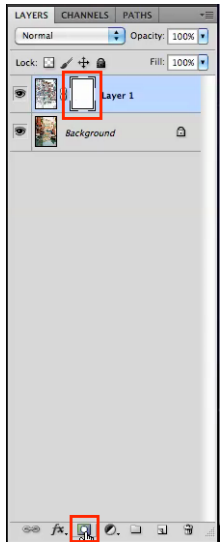


figure 7. Cropped screenshots for (a) applying the "Find Edges" style and (b) Clicking to make a layer invisible.

description and a screenshot of each step, along with a short video clip demonstrating the command action. For example, a screenshot of a dialog box enhances the instruction "*Align two or more selected layers based on their pixel content*," and its video clip shows continuous mouse action from moving to the menu, expanding the submenu, clicking on the feature, adjusting the parameters in the dialog box, and clicking the OK button to perform the action. MixT includes the following three components (shown in figure 5):

Tutorial capturing based on editing actions

To build a structured tutorial, there are two main approaches: help tutorial authors create new tutorials, or transform existing tutorials into step-by-step videos. Our system adopts the former approach by extending previous work that generates instructions in text and images [3] and combining with screen capture video demonstration. We capture and determine the time when an editing *action* is performed by the author and use these timestamps to correlate and divide the video tutorial into action-based steps. The Tutorial Builder[1], an Adobe Photoshop plug-in, enables us to record each

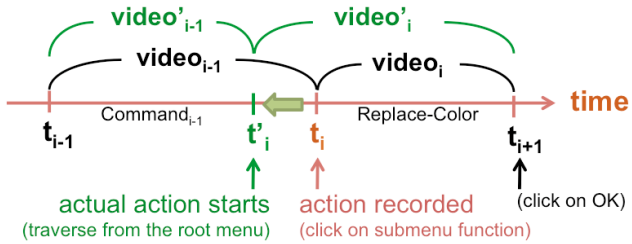


figure 6. Finding the starting time of a command triggered by traversing the UI menu in a video clip

1 Adobe. Tutorial Builder. Available at: http://labs.adobe.com/technologies/tutorialbuilder/

user command, such as opening a file ("File-Open"), selecting a region ("LassoTool"), and hiding a layer ("HideLayer"), and to generate step-by-step tutorials with textual instructions. Based on this tool, we developed an action tracker to record commands and their timestamps. We then map these data to the video screen recording as the starting time of each action to generate video segments, specific to each step.

Video segment boundary adjustment

Such an action-based approach, however, fails in certain situations because the timing of the recorded action does not always coincide perfectly with the UI interactions. One typical example is a menu navigation task (from the root menu *Image* > *Adjustment* > *Replace Color*) that invokes a dialogue box in which the user can adjust parameters (e.g. the fuzziness of a selection using a slider): the action timestamp is recorded only when the user clicks "Replace Color" in the submenu, which happens *after* the mouse hovering action starts from the root menu (figure 6).

Therefore, we include a Photoshop interface model of the menu hierarchy to acquire information about the menu path to trigger a command. Given the menu path of a particular tutorial step, we apply a computer vision technique, *template matching* [2] with an image corpus of Photoshop's menu panels to identify the key frames in the video clip: 1) the beginning frame of the action with the root menu "Image" that moves forward the segment boundary, and 2) the frame with the most matching expanded panels (i.e. when all the relevant panels are visible, including "Adjustment" and "Replace Color"), which we then crop and use as the best screenshot (figure 7a).

Mixed tutorial interface

Finally, we present visual information based on the commands of a tutorial workflow as a web page for users to navigate. The content includes: text instructions (generated from the Tutorial Builder), cropped screenshot images (highlighted manually by tutorial authors as the red boxes in figure 7b, but such annotations can be automated based on recorded operations [3]), and video clips (dynamically controlled using YouTube video player, which provides a familiar interface for users to navigate the timeline). We also include additional features to allow users to 1) replay each video clip and 2) drag-select several steps to play a longer video in case they are interested in a sequential actions such as clicking on a quick selection tool, selecting, revising the selection, and creating a mask. However, we believe these sequential steps should be automatically combined based on the granularity, which is an important challenge that requires detailed analysis of the large set of commands available in the software application. We address this problem in our on-going work by determining command relations in the menu hierarchy and similar features.

Conclusion and Future Work

We presented MixT, a system that automatically generates web tutorials with text, images, and videos. This design was based on our formative user study, which demonstrated that a mixed presentation helps users to avoid missed steps. To understand if this new type of tutorial could help users retain new skills, and to evaluate the correlation between command types and users' preferred information types, we will conduct a larger scale user study with more complicated tasks in the near future. We are also designing a newer version that improves the video navigation and includes after images of each step. All in all, we aim to provide a more supportive tutorial environment, for both tutorial authors and application users.

References

[1] Berthouzoz, F., Li, W., Dontcheva, M., & Agrawala, M. A Framework for Content-Adaptive Photo Manipulation Macros: Application to Face, Landscape and Global Manipulations. In *Proc. TOG 2011*, ACM Press (2011).

[2] Brunelli, R. Template Matching Techniques in Computer Vision: Theory and Practice. Wiley (2009)

[3] Grabler, F., Agrawala, M., Li, W., Dontcheva, M., & Igarashi, T. Generating Photo Manipulation Tutorials by Demonstration. In *Proc. SIGGRAPH '09*, ACM Press (2009).

[4] Grossman, T., & Fitzmaurice, G. ToolClips: an Investigation of Contextual Video Assistance for Functionality Understanding. In *Proc. CHI '10*, ACM Press (2010).

[5] Grossman, T., Matejka, J., & Fitzmaurice, G. Chronicle: Capture, Exploration, and Playback of Document Workflow Histories. In *Proc. UIST '10*, ACM Press (2010).

[6] Palmiter, S. and Elkerton, J. Animated Demonstrations vs Written Instructions for Learning Procedural Tasks: a Preliminary Investigation. In International Journal of Man-Machine Studies (1991), 34, pp. 687-701.

[7] Palmiter, S. and Elkerton, J. Animated Demonstrations for Learning Procedural Computer-Based Tasks. Human-Computer Interaction. 8, 3 (1993), pp. 193–216.

[8] Pongnumkul, S., Dontcheva, M., Li, W., Wang, J., Bourdev, L., Avidan, S., & Cohen, M. F. Pause-and-Play: Automatically Linking Screencast Video Tutorials with Applications. In *Proc. UIST '11*, ACM Press (2011).