

Active Capture: Integrating Human-Computer Interaction and Computer Vision/Audition to Automate Media Capture

Bibliographic Reference:

Marc Davis. "Active Capture: Integrating Human-Computer Interaction and Computer Vision/Audition to Automate Media Capture." In: *Proceedings of IEEE International Conference on Multimedia and Expo (ICME 2003) Special Session on Moving from Features to Semantics using Computational Media Aesthetics in Baltimore, Maryland*, IEEE Computer Society Press, Vol. II, 185–188, 2003.

ACTIVE CAPTURE: INTEGRATING HUMAN-COMPUTER INTERACTION AND COMPUTER VISION/AUDITION TO AUTOMATE MEDIA CAPTURE

Marc Davis

University of California at Berkeley
School of Information Management and Systems
<http://garage.sims.berkeley.edu>
marc@sims.berkeley.edu

ABSTRACT

While the devices for media capture have advanced from mechanical to computational since the invention of photography and motion pictures in the 19th century, their underlying user interaction paradigms have remained largely unchanged. Current interaction techniques for media capture do not leverage computation to solve key problems: the skill required to capture high quality media assets; the effort required to select useable assets from captured assets; and the lack of metadata describing the content and structure of media assets that could enable them to be retrieved and (re)used.

We describe a new interaction and processing paradigm for media capture that redefines capture as a control process with feedback. By integrating human-computer interaction and computer vision and audition into an “Active Capture” process, we overcome the limitations of current media capture devices, algorithms, and interaction techniques. Active Capture leverages media production knowledge to automate direction and cinematography and thus enables the automated production of annotated, high quality, reusable media assets.

1. INTRODUCTION

Since the invention of photography and motion pictures in the 19th century, the apparatus for capturing still photos and moving pictures has been subject to continual invention and refinement—the interaction paradigms for media capture have not. Photographers and videographers today face a recurrent set of challenges in trying to capture high quality, reusable media assets. These problems are especially vexing for consumers who lack the time, resources, and expertise of media capture professionals. These challenges involve having the skill and taking the time to:

- Ensure image and/or sound quality at capture time (framing, lighting, desired capture content, etc.)

- Find and select desired captures from the set of captured assets (this problem is especially difficult for time-based media such as video)
- Process and edit media assets after capture (made especially difficult by the lack of metadata describing the content and structure of media assets)

Research that combines media production knowledge, human-computer interaction design, and automated media analysis can address these challenges by fundamentally rethinking and reinventing the media capture process, rather than merely trying to optimize the process within its current device, algorithm, and interaction paradigms.

2. FROM OLD TO NEW CAPTURE PARADIGMS

When consumers produce digital media today, they engage in a capture and production process that involves numerous and often difficult manual steps (see Figure 1).

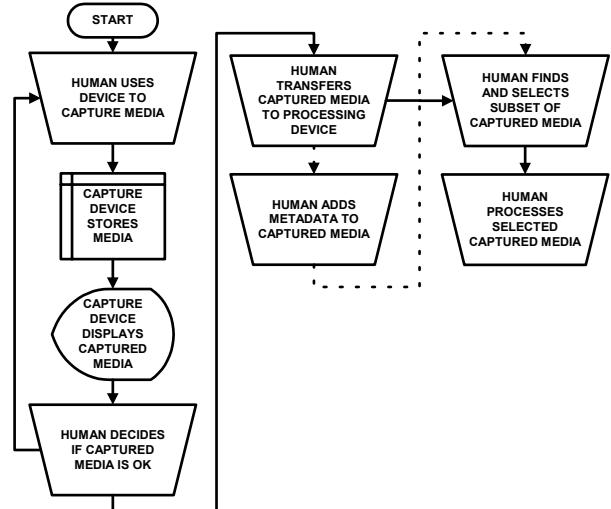


Figure 1: Current Media Capture and Production Process
(Dotted lines indicate optional steps)

Starting with the media capture process, users must perform a variety of tasks: taking a photograph or shoot-

ing video; assessing the quality/sufficiency of the captured media, and if found insufficient, beginning the capture process again; transferring the captured media to a storage and processing device (usually a computer). Then only in exceptionally rare cases (usually professional archivists or devoted hobbyists) does a user provide metadata to describe the content and structure of the captured media. When the captured media is to be processed (edited, printed, shared, etc.), the user often has to browse through a large quantity of captured media assets to find and select the desired content (this selection process is especially onerous with video data).

One might ask, what is wrong with this process? People have been taking photographs and shooting video for years, why change the media capture process? The reason is that the media capture process is both the beginning of the notoriously difficult media editing process and the key to its automation. Without editing, most consumer video and photography suffer from poor production values (think how often you look forward to seeing your friends' vacation videos?). Researchers have for years attempted to reinvent the media editing process to solve the production problems of home video and photography [1-6]. With a few exceptions [7-9], what previous researchers have neglected to do is to approach the media production problem at its source—*at the point of capture*.

Automation of the media capture process can be accomplished by *inverting* the current media capture paradigm (see Figure 2).

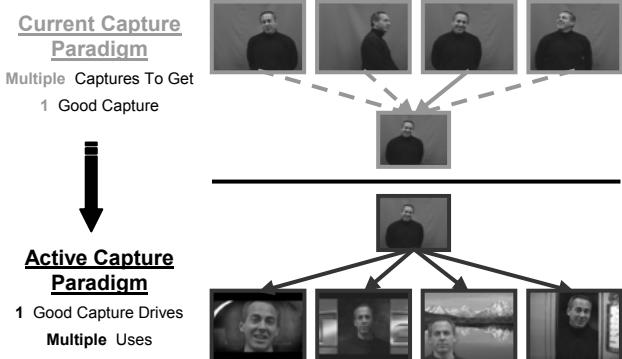


Figure 2: Inverting the Current Media Capture Paradigm

In this new media capture paradigm, which we call “Active Capture,” traditionally manual, post-capture tasks are accomplished through automated means *during the capture process*. The inversion of the media capture paradigm takes place not only in the automation of manual post-capture tasks at capture time, but also in the quality and quantity of media captures required to meet users’ goals. In the current media capture paradigm, users capture more media than they need in to order to ensure they have captured the media they want. This is especially true for video, but also applies to photography. Rather than

taking numerous photographs to get one good one or shooting hours of video to extract a few memorable minutes, in Active Capture the capture device works with the user in an automated process to capture a smaller number of high quality, annotated media assets that can be automatically used and reused in a variety of contexts.

3. ACTIVE CAPTURE

In schools and shopping malls across the nation, each year millions of children have their portraits taken by professional photographers who use a variety of techniques to coddle, entreat, and compel them to look at the camera and smile. In professional motion picture production, directors instruct actors, record them, provide feedback, and reshoot until they “get the shot.” In consumer photo and video capture, amateur photographers, videographers, and their subjects are all familiar with the interaction technique of asking subjects to “Say Cheese!” in order to improve the likelihood that the resulting photo or video will feature the subjects smiling and looking at the camera. The central idea in these media capture examples and in the Active Capture paradigm is engaging the user in a *control process with feedback* to iteratively capture until a satisfactory result can be achieved or a timeout is required (see Figure 3).

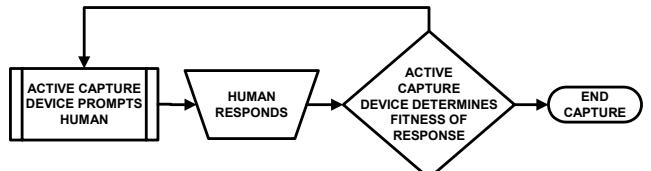


Figure 3: Active Capture Control Process with Feedback

In portrait photography, motion picture direction, and “say cheese” interactions, humans both prompt and evaluate the responses of the capture subjects. In Active Capture both prompting and response evaluation can be achieved by the media capture device itself. An Active Capture device uses audio and/or visual cues to prompt the capture subject to perform some desired action (e.g., smiling and looking at the camera). Through real-time audio and video analysis an Active Capture device determines the fitness of the subject’s response in relation to some predetermined capture parameters. If the capture satisfies these parameters, the capture process is complete. If not, the Active Capture device prompts the user again (using clarifying instructions) until a suitable response is achieved or the process has timed-out.

Figure 4 illustrates an example of Active Capture’s control process with feedback. This example depicts the Active Capture routine for capturing a high quality and highly reusable shot of a user screaming. Based on our work in media automation [1, 10, 11], we have developed automatic media production systems that can use a scream

shot (as well as others) in a variety of contexts (e.g., MCI commercial, 7Up commercial, Godzilla movie scene, banner ads, Flash animations, Blair Witch movie trailer, etc.). In order to capture a shot of the user screaming, the system prompts the user to look at the camera and scream. The system has a minimal average loudness and overall duration it is looking for, and like a human director can prompt the user accordingly (e.g., scream louder, scream longer) in order to capture a loud enough and long enough scream shot. This simple example is meant to illustrate the basic concept of the Active Capture interaction paradigm: a control process with feedback.

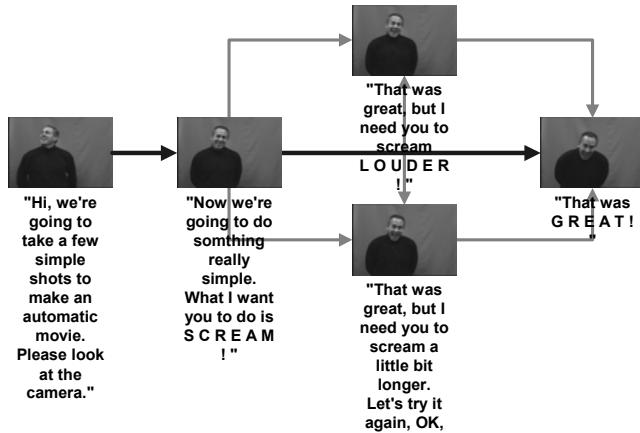


Figure 4: Active Capture Process for Capturing “Scream” Shot
(Quotes are verbal instructions from the Active Capture device.
The darker arrows represent an error-free path. The lighter arrows are
error correction loops.)

We have developed more sophisticated Active Capture routines for “look at the camera” and “head turn” shots that involve real-time video analysis and feedback for a wider variety of error conditions (not looking at the camera, not looking away from the camera, not standing still, moving too slow or too fast). Part of the larger research agenda for Active Capture is to investigate the sweet spot among the types of shots (especially human actions) that we can: easily elicit people to produce; reliably parse; and afford the greatest degree of reusability in the creation of personalized and customized media assets.

3.1. Integrating capture, processing, and interaction

The Active Capture paradigm reinvents the media capture and production process by integrating three elements that have familiar pair-wise combinations, but until now have not been integrated into an automated media capture system: *capture*, *processing*, and *interaction* (see Figure 5). Active Capture combines these three processes to enable the communication and interaction among the capture device, human agent(s), and the environment shared by the device and agents by integrating technology and tech-

niques from computer vision and audition, human-computer interaction design, and media production practice (direction and cinematography). Unlike previous systems focused on meeting and lecture capture that have endeavored to automate cinematography and editing [12], the Active Capture paradigm also automates direction, in addition to automating cinematography and editing, in order to interactively affect and help shape the creation of the captured content.

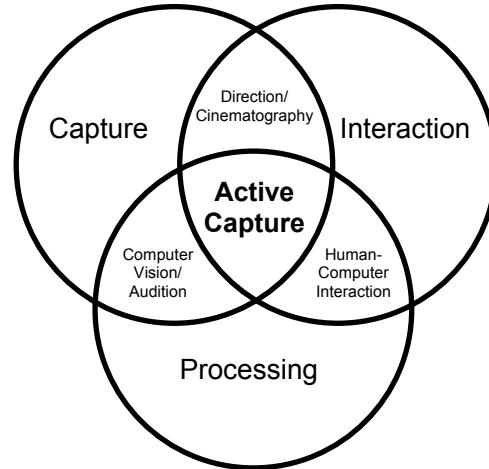


Figure 5: Active Capture Integrates Capture, Processing, and Interaction

With Active Capture, we bridge the “semantic gap” [14] by integrating human-computer interaction and computer vision/audition at the point of media capture. We overcome the limitations of standard computer vision and audition techniques by using human-computer interaction design to simplify the world and the actions that the vision (and audition) algorithms need to parse. As a result, we can use very simple, robust analysis algorithms coupled with judicious interaction design in an interactively simplified parsing context. By using HCI to reinvent media capture as a “human-in-the-loop” algorithmic process, we believe we also indicate a new and fruitful direction for multimedia researchers.

3.2. Interaction modes for Active Capture

In our research we have identified four distinct interaction modes for Active Capture:

- *Directed Performance*
The user is directed to perform a specific action or utterance (e.g., “scream”).
- *Improvised Performance*
The user is directed to improvise an action or utterance. This mode supports a spectrum of generality from very specific actions and utterances to more general ones (e.g., from “scream in abject terror” to “show an intense emotional reaction”).

- *Record Structured Activity*
The user is recorded while engaged in an activity whose structure the system knows enough about to be able to parse and process it automatically.
- *Agit Prop*
The system elicits the user's response to an unexpected stimulus (e.g., system yells "Boo!" => user utters a startled scream).

We have implemented examples of Directed Performance (See Figure 4). The directorial quality of Directed Performance user instructions can range from explicit commands to fairly unobtrusive suggestions during the capture process. We have also built Active Capture prototypes that use the Record Structured Activity interaction mode to capture high-quality, annotated reusable media assets. Agit Prop has a long history of success from hand-buzzers to amusement park fun-houses in eliciting predictable user reactions to stimuli. The experience design of Active Capture interactions and devices will productively draw on a variety of sources for inspiration ranging from human-computer interaction design, consumer electronics interfaces, motion picture production, theater and improvisation techniques, and theme park attraction design.

4. FROM ACTIVE CAPTURE TO REUSABLE MEDIA ASSETS

By guiding the capture process, Active Capture provides rich and reliable metadata at the beginning of the media production process. Active Capture redefines the steps and agents involved in media capture and production to re-envision it as an automated process (see Figure 6) [11].

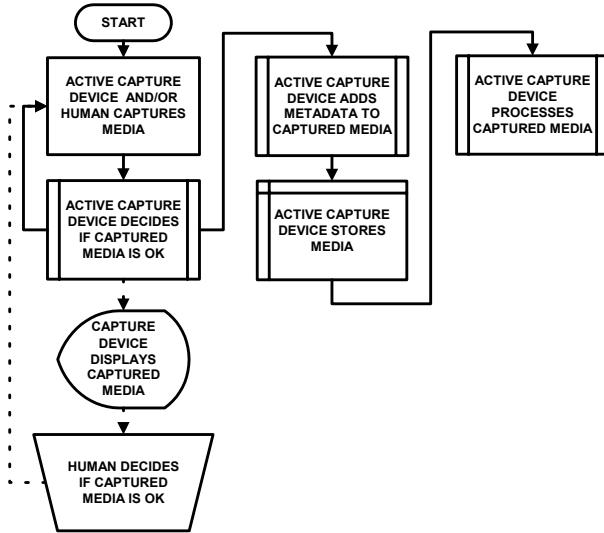


Figure 6: Active Capture Media Production Process

By producing annotated media assets, the Active Capture process becomes the first and key step in a new media production process that leverages media metadata

and knowledge about cinematic structures and functions to automatically produce high quality personalized and customized media content from reusable components [1, 8, 10, 11, 13].

5. CONCLUSIONS

By integrating capture, processing, and interaction into a control process with feedback, Active Capture overcomes many of the limitations and difficulties inherent in the traditional, manual media capture process. As a result, media assets can be captured in which quality is automatically ensured at capture time. Most importantly, the Active Capture process automatically produces rich and reliable metadata that can be used not only to make it easier to find and select media assets, but to automate the entire process of media production and reuse.

6. REFERENCES

- [1] M. Davis, "Media Streams: An Iconic Visual Language for Video Representation," in *Readings in Human-Computer Interaction: Toward the Year 2000*, R. M. Baecker, J. Grudin, W. A. S. Buxton, and S. Greenberg, Eds., 2nd ed. San Francisco: Morgan Kaufmann Publishers, Inc., 1995, pp. 854-866.
- [2] A. Bruckman, "The Electronic Scrapbook: Towards an Intelligent Home-Video Editing System," in *Media Laboratory*. Cambridge, Massachusetts: MIT, 1991.
- [3] A. Kuchinsky, C. Pering, M. L. Creech, D. Freeze, B. Serra, and J. Gwizdka, "FotoFile: A Consumer Multimedia Organization and Retrieval System," presented at CHI '99, Pittsburgh, Pennsylvania, 1999.
- [4] A. Grgensohn, J. Boreczky, P. Chiu, J. Doherty, J. Foote, G. Golovchinsky, S. Uchihashi, and L. Wilcox, "A Semi-automatic Approach to Home Video Editing," presented at UIST '00, San Diego, California, 2000.
- [5] B. Shneiderman and H. Kang, "Direct Annotation: A Drag-and-Drop Strategy for Labeling Photos," presented at International Conference Information Visualisation (IV2000), London, England, 2000.
- [6] J. Casares, B. A. Myers, A. C. Long, R. Bhatnagar, S. M. Stevens, L. Dabbish, D. Yocom, and A. Corbett, "Simplifying Video Editing Using Metadata," presented at Designing Interactive Systems (DIS 2002), London, England, 2002.
- [7] G. Davenport, T. G. Aguirre-Smith, and N. Pincever, "Cinematic Primitives for Multimedia," *IEEE Computer Graphics and Applications*, vol. 11, no. 4, pp. 67-75, 1991.
- [8] F. Nack and W. Putz, "Designing Annotation Before It's Needed," presented at MM '01, Ottawa, Canada, 2001.
- [9] S.-F. Chang, "The Holy Grail of Content-Based Media Analysis," *IEEE MultiMedia*, vol. 9, no. 2, pp. 6-10, 2002.
- [10] M. Davis and D. Levitt, "Time-Based Media Processing System (US Patent 6,243,087)." USA: Interval Research Corporation, 2001.
- [11] M. Davis, "Editing Out Video Editing," *IEEE MultiMedia*, vol. 10, no. 2, pp. 2-12, 2003.
- [12] Q. Liu, Y. Rui, A. Gupta, and J. Cadiz, "Automating Camera Management for Lecture Room Environments," presented at SIGCHI'01, Seattle, Washington, 2001.
- [13] F. Nack, "The Future of Media Computing," in *Media Computing: Computational Media Aesthetics, The Kluwer International Series on Video Computing*, C. Dorai and S. Venkatesh, Eds. Boston: Kluwer Academic Publishers, 2002, pp. 159-196.
- [14] C. Dorai and S. Venkatesh, "Computational Media Aesthetics: Finding Meaning Beautiful," *IEEE Multimedia*, vol. 8, no. 4, pp. 10-12, 2001.