Integrating Contextual Video Annotation into Media Authoring for Video Podcasting and Digital Medical Records

I-Jong Lin, Hui Chao
Digital Printing and Imaging Laboratory
HP Laboratories Palo Alto
HPL-2007-9
January 22, 2007*

video annotation, multimedia processing, digital medical records, podcasting

In this paper, we demonstrate how rich media annotation can enable two new applications for video podcasting and digital medical records. At WIAMIS 2004, we introduced an innovative video annotation technology called Active Shadows, that captures a virtual presence interacting with a displayed image and overlays on top of a digital image. With Active Shadows, we combine the expert opinion and presence (of a professor and doctor) with the informationally rich images (presentation slides and medical images, respectively). We identify current forms of media whose value can be enhanced through this type of annotation, and introduce a basic workflow to produce new forms of media. This paper specifically discusses two experimental media workflows, one for authoring video for portable video devices (video podcasting) and another for integrating diagnoses with medical imaging for digital patient records.

* Internal Accession Date Only
Approved for External Publication

Integrating Contextual Video Annotation into Media Authoring for Video Podcasting and Digital Medical Records

I-Jong Lin { i-jong.lin@hp.com }, Hui Chao {hui.chao@hp.com}
Hewlett Packard Laboratories, Palo Alto, CA, USA

December 1, 2005

In this paper, we demonstrate how rich media annotation can enable two new applications for video podcasting and digital medical records. At WIAMIS 2004, we introduced an innovative video annotation technology called Active Shadows [1], that captures a virtual presence interacting with a displayed image and overlays on top of an digital image. With Active Shadows, we combine the expert opinion and presence (of a professor and doctor) with the informationally rich images (presentation slides and medical images, respectively). We identify current forms of media whose value can be enhanced through this type of annotation, and introduce a basic workflow to produce new forms of media. This paper specifically discusses two experimental media workflows, one for authoring video for portable video devices (video podcasting) and another for integrating diagnoses with medical imaging for digital patient records.

1 Introduction

Already, the two trends of increasing digital connectivity and increasing ubiquity of cheaper digital cameras empowers individuals to easily distribute and produce their own digital content. Broadband connections that have the capacity for stream video are commonplace; megapixel cameras are included on cellphones. However, the massive amounts of available video through a distributed sensor network will not organically create content.

We feel the next step is evolution of media is to enable a person to add value to the consumed content. The audience will play a role in how content is created, blurring lines between production and consumption. This evolution requires not only a physical transmission line and connected imaging devices, but also an authoring construct so that a person add relevant content that fits within context of the media.

We define *contextual annotation* as the process of capturing the explanatory annotation (audio, video or both) within the context of digital media. We can divide the annotation in two parts: 1. the link into the media that contextualize the recording, and 2. the recording itself. Starting with structured content, we build up new content by iteratively layer contextual annotation. By allowing annotation over the network, we create new type of content via collaborative authoring paradigm.

In the domain of text-based media, we already have successful examples of this type of collaborative authoring. For programming source code, Open Source projects such as the Linux kernel and Apache HTTP server creates free software products that rival or beat their commercial counterparts in quality and stability. For syndicated content, blogs and wikis are directly competing with newspaper columns. For reference content, Wikipedia is working collaborative encyclopedia with over 750,000 articles in English alone. Contextual text annotation are the simple operations of text insertion, deletion and commenting. However, contextual video annotation of images and video require more specialized technology.

In this paper, we demonstrate how to leverage the power and the reach of the current network infras-

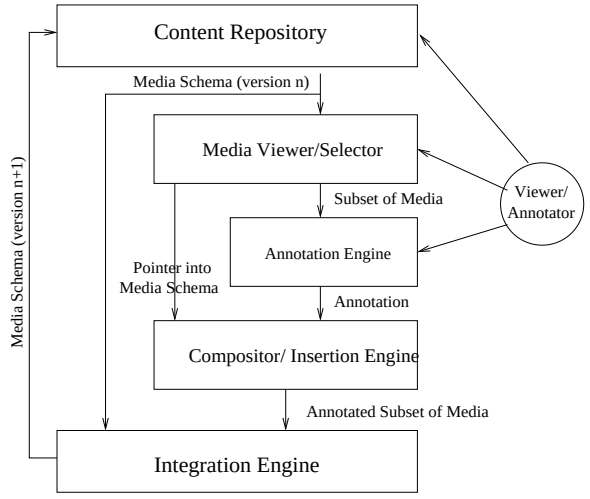

Figure 1: Contextual Annotation workflow

tructure into the rich media authoring process via a video annotation tool. We will describe the generic workflow and then discuss map the contextual video annotation workflow that uses Active Shadows. We then apply this workflow to different types of digital media to create new forms of digital media.

2 Contextual Annotation

As shown in Figure 1, the flowchart for contextual annotation within a iterative version process is a five step process. First, an author selects the content from the repository via index or search engine. Second, while consuming the content and understanding the overall structure of the media, the authors select the relevant subset of the content to annotate. Third, the author annotes this subset with his/her own content. Fourth, a link is created to represent the selection of content and is paired with the annotation; the annotation is transcoded to match the select content. Fifth, the annotation and link are encoded back into the content itself and the content in repository is updated.

We believe the most effective workflow requires a method of annotation where the content and annotation are of similar modality: the datatype of annotation and content are similar and the viewing environment is similar to annotation environment. This constraint has two benefits: 1) simplification of the annotation infrastructure, minimizing the number of codecs required across distributed authoring environment, and 2) ability to the seamlessly switch from browsing to authoring environment, removing extra training and hardware for annotation.

Most importantly, we require technologies that can maintain context while moving the author from media consumption to annotation. For text-based media, the display, computer, keyboard and mouse provide a sufficient interface for both browsing and annotating media; programmers around the world communicate in ASCII plain text. However, for richer forms of media such as video and images, we require more complex technologies in order to avoid switching modalities.

3 Active Shadows Annotation

In this paper, we use the Active Shadows algorithm as our primary means annotate video. Active Shadows captures the person naturally interacting with a displayed image or video and produces video that maintains the physical-digital correspondence, i.e. the location in the image that you are pointing on the physical screen, is the same pixels that your digitized presence of yourself is pointing to in the video.

The Active Shadows algorithm produces low latency (under 200ms) and high quality video object segmentation of a person, occluding a computer-controlled display of any type (LCD, projector, CRT, plasma) in front of video camera device. This algorithm sets up a causal video feedback loop that can resolve ambiguous visual occlusions by adaptively modifying the displayed image in real-time. These real-time modifications to the display manifest themselves as if the camera were a virtual light source and was casting a reverse shadow onto the display. Active Shadows gives the same output as a chromakey system except that the user is physically interacting with the displayed image, instead of a colored background. With this setup and a wireless microphone, the system produces segmented video at ap-

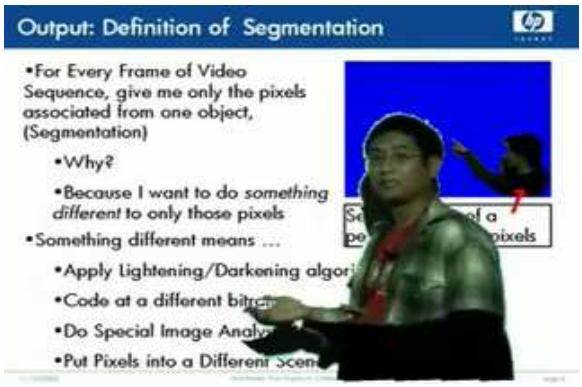

Figure 2: A researcher gives an explanation of Active Shadows algorithm through video created by the Active Shadows algorithm.

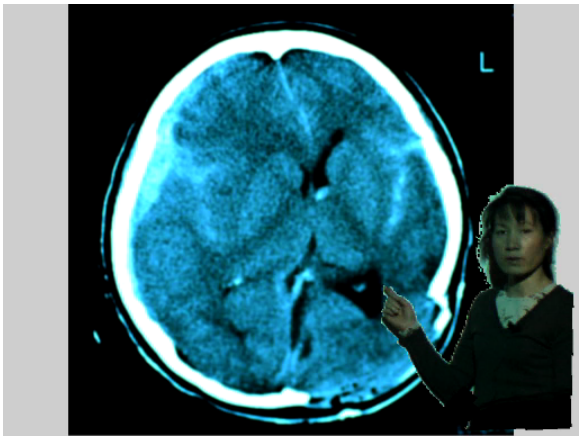

Figure 3: The annotator is pointing at right to left midline shift in a head CT (Computer Tomography) scan image.

proximately 7 fps and seamlessly composites presentation slides and segmented video of the speaker to create a multi-layered video representations with synchronized audio of the speaker.

Active Shadows immerses the viewer in a image or video, and simultaneously captures the viewers experience. Building off of this technology, we can create two applications for rich media authoring workflow: 1) video podcasting and 2) digital electronic patient records.

4 Applications

In these two applications, we approach the authoring of new forms of media as the process of adding value via rich media annotation to well-known forms of media. The first overlays video of a presenter on top of a slidesets to produce video content for portable video devices; the second pairs complex medical imaging data and pairing with opinions from medical professionals.

Video Podcasting Our first application is the production of educational content for a small handheld video viewing device with full video framerate, but limited resolution (such as the video iPod with 320 by 240 resolution). We consider portable video players as an important vehicle for educational content. However, due to the small screen size and limited resolution of these low power devices, we need to balance our content between the background image and richness of the video overlay. Standard presentation slidesets can be used as background images after their content is filtered to remove objects and graphics that are illegible at lower (320x240) resolution. Furthermore, the slidesets must be divided into episodes so that the viewing time for each episode is about 7-20 minutes. A full hour-long presentation will usually run 7-9 episodes. In figure 2, we show the frame of the second episode. This video content is currently available off of two sites: 1) my public website at Hewlett-Packard Labs (`http://www.hpl.hp.com/personal/I-Jong_Lin/`) and 2) the iTunes directory hosted by Apple Computer through their iTunes applications (under "Hewlett Packard Experimental Video Podcasts").

This video content is currently available off of two sites: 1) my public website at Hewlett-Packard Labs (`http://www.hpl.hp.com/personal/I-Jong_Lin/`) and 2) the iTunes directory hosted by Apple Computer through their iTunes applications (under "Hewlett Packard Experimental Video Podcasts").

Digital Patient Records Based on the Active Shadows, CMAS (collaborative medical annotation system) is a rich media authoring tool for recording diagnostic result based on medical images and videos. We can intuitively capture physical annotation of displayed digital images/video while recording the synchronized audio. Physical annotation can be as simple as a laser pointer spot to full-motion overlaid video. The diagnostic results based on digital radiological images, videos and digital pathological slides will become metadata of those medical images and video to video and image annotation. In the case of video input, key frames are identified and annotation is connected to one or more key frames of the video. A physician points to and marks up the areas of interest using a laser pointer or a markup pen when giving his or her medical opinion to the background image. The motions of the marking device, the synchronized audio or even the presence of the physician are recoded as the annotation of input image or video. The trajectory of the marking device can be recorded and displayed as a visible highlight in image or The annotation can take different forms: a visible highlight, scalable vector graphics(SVG), a second layer of overlaid video, or a metadata link of the previous data. The DICOM-X architecture [2] explored the possibility of integrating medical images into an XML based Electronic Health Records; we could expand this system to accept annotations of image, video and audio. Annotation of radiological images, echocardiogram, procedure videos from doctor captures not only facts and measurements of the patient diagnostic result , but also the development of the diagnostic medical opinion. Currently, more medical diagnoses are done through voice recording [4]. Seamlessly incorporation of richer media into the medical record system could further reduce the process-related error and improve the accuracy and integrity of the data [3]. With the advancement in medical imaging, medical procedure and image guided surgery, it becomes more important to accurately capture as much diagnostic information as possible while maintaining the context of the patient's overall health.

5 Conclusion

In conclusion, this paper shows how the video annotation can become an integral part of video authoring workflow, enabling new types of applications with new forms of media. From a generic workflow that is applicable to Open Source coding projects such as Linux or any numerous projects hosted by Sourceforge, we apply the video annotation system called Active Shadows that allows for immersive viewing of images and video while simultaneously capturing the video annotation in the form of the person's physical interaction with a projected digital image. Both of the applications that we have presented can also work within a real-time collaboration infrastructure, allowing for real-time feedback on annotation and enabling conversations with the digital media.

References

[1] I-J. Lin. Active shadows: Real-time video object segmentation in a camera-display space. In *International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2004), Lisboa, Portugal*, 2004.

[2] B. Jung. Dicom-x seamless integration of medical images into the her. In *Proceedings of the IEEEE Symposium on Computer-Based Medical Systems*, 2005.

[3] T. Welzer, B. Brumen, I. Golub, and M. Druzovec. Medical diagnostic and data quality. In *Proceedings of the 15th IEEE Symposium on Computer Based Medical System, (CBMS 2002)*, 2002.

[4] J.A. Hollbrook. Generating medical documentation through voice input: The emergency room. *Topics in Health Records Management*, 12(3):58–63, 1992.