CONTINUOUS RESPONSE EVALUATION OF DIGITAL VIDEO CLIPS OVER THE INTERNET

Griff Richards, Tony Tin & Hongxing (Bill) Geng

Athabasca University
Athabasca, Alberta, Canada
{griff|tonyt|billg}@athabascau.ca

Abstract

Continuous response evaluation of video and film has been a useful method in media. It has been particularly prevalent in advertising as it allows producers to pinpoint events that evoke particularly strong audience responses. When searching for a way to evaluate the effectiveness of digital video clips (podcasts) a continuous response evaluation system was developed and deployed over the internet. This paper discusses the preliminary results of testing the method with adult and high school learners to evaluate a series of video podcasts on meta-cognitive success strategies. When fully deployed this tool will enable the real time viewer evaluation of internet video formats such as Youtube. The resulting data could be used to rate the overall interest of video programs or index specific scenes for educational or entertainment contexts.

Keywords - Video. vidcast, evaluation, internet

1 INTRODUCTION

Broadband internet access has enabled the widespread transmission of digital video clips over the internet. Indeed, Alexa.com, the provider of internet usage metrics rates Youtube.com as the third most popular internet site world wide [1] behind Google and Yahoo. Youtube,com does provide a simple five-star rating system by which any viewer can rate a video and also leave comments about the video, however these popularity ratings provide little information about the usefulness of the content for instruction. Recently, in the context of another project [2] the lead author had produced a series of video clips for use in internet instruction and wanted to evaluate their design and the interest level of the content for on-line learners. In traditional instructional settings, it would be possible to assemble a small focus group representative of the target audience, and after a screening the video discuss its merits. Technology-based methods have also been used in preparing advertisements and instructional videos for some time. Continuous Response Measures (CRM) first appeared in the 1930's for analyzing radio shows [3, 4]. In 1980 Nickerson [5] demonstrated a CRM system driven by an Apple II microcomputer that allowed for second by second analysis of a video so that producers, advertisers and educators can determine key incidents that evoke audience reactions. Baggaley [6] reports the use of push-button data collection technology for gathering continuous audience responses when evaluating video and live events. CRM systems have become more elaborate over the years and at least one patent touts the correlation of EEG with galvanic skin response and facial expressions to elicit the true audience reaction to a video event [7]. The goal of this paper is to describe a prototype video evaluation tool that has been prototyped for simple user input CRM data collection over the internet.

2 CONTINUOUS RESPONSE EVALUATION

2.1 Continuous Response Evaluation Method

Continuous Response Evaluation involves the open solicitation of responses from a target audience throughout the presentation of a stimulus video program. Instead of booing or cheering aloud, the audience pushes buttons or turns knobs on an input device to report their approval or disapproval of the presentation. The responses and the elapsed time (or timecode) of the response are collected for statistical tabulation. While technically the sampling periods are limited to a sequence of milliseconds, for practical purposes the response is continuously open. Frequency charts can be generated to correlate the audience reactions with the timecode of the video events. If demographic information is known about the respondents, then the data can be compared between sub-groups in the sample. For example, men might respond favourably to an event which women find disagreeable.

In a controlled environment, such as a room with a focus group, the data collection process is quite simple – each participant is provided an input device, and they are asked some warm-up questions such as “If you are a male. Press A or if you are female, Press B”. Thus, demographic information ca be collected while familiarizing the participants with the input device. Typically, a short practice video will precede the target video. Researchers might also ask additional questions before or after the video screening – for example if for a political campaign, they might ask questions about attitude towards a particular party or candidate, and these questions might be repeated afterwards to gauge any shift in attitude which might be attributable to the content of the video segment, and as a measure of the criterion validity of the CRM data [6]. Such tactics are commonplace in marketing campaigns where the goal is to help design video advertisements that provoke the largest attitudinal change with the least negative response to the way the message is presented. The higher the stakes, the more formative evaluation that is likely to take place.

The goals of a formative evaluator of educational video is not unlike that of the marketer – a good educational video producer wants to determine the degree to which a video hold the attention of the audience and the extent to which their message is received by the audience. At question is the efficiency of the message design and production quality. When video programs tended to follow a 28 minute television format, there were several opportunities for a producer to embed redundant messages and show an educational point in several contexts. Today, the length of a digital video is much shorter –Youtube limits the length of a video to 10 minutes, music videos and video podcasts or “vidcasts” tend to run about 3 minutes, and advertisements either 30 or 60 seconds. Vidcasts need to be well-designed to gain attention and get their message across in only 3 minutes – anything longer creates a file size that dissuades potential downloads. With only 3 minutes, a producer needs to ensure that the message is transmitted efficiently i.e. in a way that captures and maintains the audience’s attention, in a manner that is memorable, and in a very short period of time.

To investigate the potential of continuous response evaluation system for internet vidcasts, the authors developed and tested a prototype system.

2.2 Design of the Continuous Response Evaluation Internet Prototype

The Continuous Response Evaluation System (CRES) is built using the following technologies: Flash, ActionScript 3.0, HTML, PHP, and MySQL as a database server. Figure-1 illustrates the system architecture.

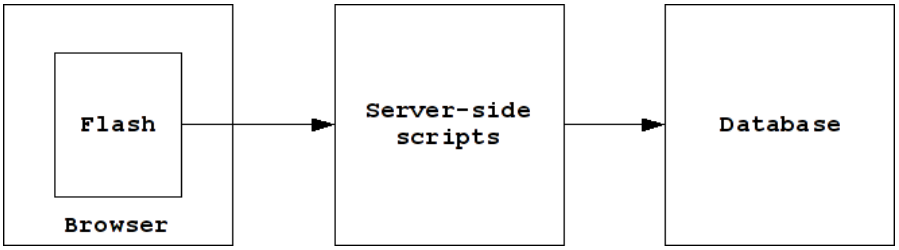

Figure 1. System Architecture

As depicted in Figure-1, flash video are embedded in an HTML window. Users access the videos through their web browser. During the play time of the videos, users are provided an approach to evaluate the videos by clicking any number of buttons that can be labelled “Like”, “Dislike” etcetera. The system collects user input and sends the data to the server-side scripts, which, in turn, wrap the data with video timecode, IP address and real time stamp information, and sends the data to the database.

A. Constraints of the system

In the current prototype, videos are embedded in a Flash window. Users cannot access the videos through iPhones or other mobile devices that do not support Flash. Another way to design the system is to separate the video window from the evaluation window and input targets. While this is more

flexible, it becomes more difficult to communicate data such as timecode from the video part to the evaluation part and leads to synchronization issues.

B. Video formats and encapsulation

To embed videos into Flash, in the system, we convert .mov format files into Flash video format files (file extension is .flv), which, in turn, are embedded into a SWF file dynamically. The SWF file can be enclosed in HTML files and runs inside. One advantage of this approach is that the largest source of internet videos, Youtube.com, also uses the Flash format.

C. Data transmission and collation

In addition to multiple videos, the continuous response evaluation prototype has questionnaires and text entry questions to allow users to respond. The current questionnaires have three parts: pre-video questionnaire, video-follow-up questionnaire, and post-video questionnaire. Before finishing a questionnaire or a video, user input is stored in the memory of client-side computers. Upon completing a questionnaire or video, the collected data will be sent to the server-side scripts for persistent storage. This messaging eliminates the constant polling that would needlessly consume bandwidth if a large number of participants at a single site were engaged in formative evaluation activities.

An extract of a typical data stream appears in Figure 2. The general format of the data is

IP address-datetime-[session ID:type:data]

The **data** format will be different for each **type** of questions. The **type** can be one of the following five types:

Q: means pre-video questionnaire

Vn: means the nth video

SQn: means the nth video corresponding questionnaire

VQ: means post-video questionnaire

Cn: means the nth text entry question

The data format of type **Q**, **SQn**, and **VQ** is `question number-option number`. Each data pair is separated by a semicolon.

In the Like vs Dislike scenario the data format of type **Vn** is `L or D @timecode`. Each data pair is separated by a comma. The data format of type **Cn** is user input text strings.

```
70.54.141.108-Thu Apr 9 21:23:26 GMT-0400 2009-[27453:Q:1-9;2-
1;3-6;4-6;5-2;6-1;7-5;8-5;9-1;10-3;11-1;12-1;13-1;14-3;15-5;16-5;17-1]
70.54.141.108-Thu Apr 9 21:23:26 GMT-0400 2009-
[27453:V1:L@63.999,L@64.73,L@64.913,L@65.461,L@67.891,L@68
.074,L@98.637,L@99.133,L@99.525,L@99.891,L@100.361,L@100.6
75,L@101.406,L@133.171,L@133.51,L@135.13,L@181.994,L@183.
535,L@194.663,L@200.515,D@207.724,D@208.404,D@209.866,D@
211.982,D@213.393,D@214.595,D@229.928,D@230.973,D@232.04
4,L@233.455,L@233.455]
70.54.141.108-Thu Apr 9 21:23:26 GMT-0400 2009-[27453:SQ1:1-2;2-
3;3-2;4-2]
```

Figure 2. Excerpt of data stream showing question responses and continuous response time codes in seconds and thousands of seconds

3 PILOT TEST

3.1 Pilot Test Method

26 adult on-line learners in a graduate course on instructional design volunteered to “view some videos”. They were given the prescribed URL and responded anonymously over a two one-week periods. The system presented pre-cursor questions, the seven video clips with English sub-titles (each followed by 3 or 4 questions about the video clip content) and then the open response summary questions about their experience using the evaluation system. During the video presentation respondents could click on either a “Like” or “Dislike” in week one, or “Interesting” or “Boring” in week two.

The collated results were downloaded into MS Excel for tabulation. An excerpt of the data is provided in Table 1.

3.2 Pilot Test Results

The initial results were quite varied. All of the respondents worked from their own workstations, usually at home or office, and used Internet Explorer if on a PC or Firefox if on a Macintosh computer. However, we found that there were some design issues and that the straight linear design of the presentation had disabled re-loading of the video when transmission errors or lags occurred in the loading of each video. This was particularly aggravating for a participant in Qatar receiving streaming videos from a server in Canada. As a result, several of the results had to be discarded.

Comments from the participants indicated that it took at least one video to understand how to effectively respond to the “like” or “dislike” input method.

A. Data collection and analysis

Fig. 1 show an excerpt of the input data stream arriving at the server. Note that the IP address, a time stamp and a “Session ID” help identify the data coming from a participant. Although the identity of the participant is anonymous, it was important to distinguish each participant’s data. The Q indicates the pre-questions consisting of demographics and familiarity with the content. SQ indicates Segment Questions, and V1, indicates an input stream of mouse clicks (L=like, D=Dislike). This format makes it relatively easy to transfer the data into a spreadsheet for analysis, although automatic analysis routine should be embedded as the project matures.

Table 1 illustrates a small sample of response data for Video 1 that has been arbitrarily clustered into 30 second intervals. The resulting chart of positive (LIKE) and negative (DISLIKE) category counts by interval is charted in Fig. 2 by coding negative counts as negative integers. Again it should be stressed that with a small number of subjects, one or two particularly active participants can exert a large amount of influence on this form of summary data. The choice of time interval for depicting the analysis is up to the producer or the analyst, and with larger numbers of participants or with more responsive participants it will be practical to have smaller time divisions, to produce a chart based upon the edit points of the video segments, or to overlay the data on the actual video as previous investigators have done [4, 5, 7]. While the sample of data is sufficient to demonstrate the principle of the prototype, ideally a CRM system should elicit a higher frequency of response.

Table 1. Sample of LIKE and DISLIKE data collated into 30 second time intervals

Subject		0-29	30-59	60-89	90-119	120-149	150-179	180-END	Total
1	LIKE				2	3	2	1	8
1	DISLIKE								
2	LIKE	1			2			2	5
2	DISLIKE								
3	LIKE							1	1
3	DISLIKE								
4	LIKE	1							1
4	DISLIKE								
5	LIKE		1					1	7
5	DISLIKE		2		1	1	1		
6	LIKE					1		1	6
6	DISLIKE	1	1		1		1		
7	LIKE				1	2	1		4
7	DISLIKE								
8	LIKE								4
8	DISLIKE		1	1	1	1			
9	LIKE				1			1	7
9	DISLIKE			2		2	1		
10	LIKE		2	2	3	1	4	3	22
10	DISLIKE		1	2	4				
11	LIKE			6	6			4	16
11	DISLIKE								
		0-29	30-59	60-89	90-119	120-149	150-179	180-END	
ALL	LIKE	2	3	8	15	7	7	14	
ALL	DISLIKE	-1	-5	-5	-7	-4	-3	0	

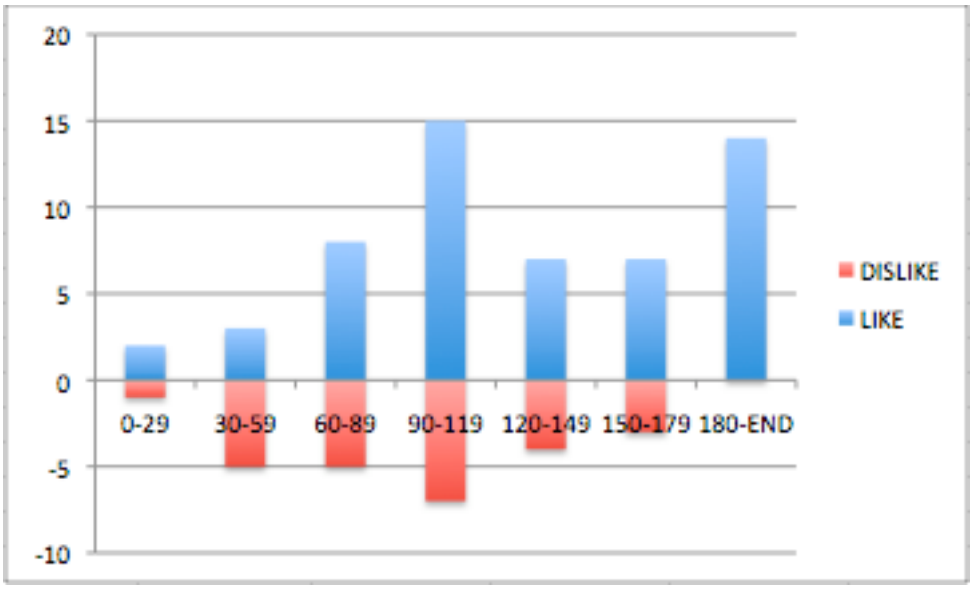

Figure 2. Chart of LIKE and DISLIKE data collated by 30 second time segments

4 DISCUSSION AND CONCLUSIONS

4.1 Reliability, Validity: Just What Is Being Measured?

While Continuous Response Measures have been used for over 80 years, they remain controversial in terms of explanatory power. Maier et al [8] review the issues of external and internal validity and note the difficulty in knowing just what a user's input indicates. They also note that continuous response measures can be compared with post-viewing questionnaires as a means of establishing criterion validity. However, trying to understand why a viewer reacts positively or negatively to a video event might only be teased out by re-viewing the video in a de-briefing activity such as a focus group, or by asking the viewer to "think aloud" during the initial screening. Maier also suggests Z-score data transformations that can be used to smooth large data sets into less noisy plots.

In our pilot test we collected open comments from the participants about the videos and about the method of evaluation. We found these same issues raised by the participants. As one put it, "How do you know what I am reacting to? Is it the bad acting, the poor script, a sloppy edit, or my own foul mood?" There is information value in knowing what viewers "like" or find "interesting", but it is difficult to delve deeper with CRM data alone.

4.2 Feedback on Interface Design

Reliability of input was also an issue – most CRM investigators stress the need for training and practice sessions with the input device before the evaluation session. Our internet pilot data showed that while some participants were very active and provided as many as 22 responses over a 230 second video clip, many more provided a mere 4 or 5 inputs and others failed to respond at all. Some of this is attributable to the primitive interface design, and some users suggested that additional feedback is required to know that the inputs are actually being received and tabulated. Some suggestions were also received about using key-presses as the on-screen movement of the cursor distracted from the message. One user commented that between listening to the video and reading the sub-titles there wasn't enough cognitive capacity to manipulate the mouse and select the click boxes. Clearly, there is room for improvement on both the interface and the pre-evaluation training procedures to ensure that subjects do indeed understand their role, the operation of the input method, and that they receive real-time feedback on their scores that hopefully will evoke additional response.

One aspect of the pilot that did work well was the collection of segment questions. The initial idea of inserting three or four multiple-choice segment questions after each video was to provide a way of knowing that the videos were actually being attended to, and that the key message was not being overlooked. The segment questions usually asked who the main character was, what was their problem and what was the key message of the video. Subjects had a high level of correct responses to the segment questions, and in future corroborating questions about the effectiveness of each video could also be added. While the segment questions can provide information about the over-all effectiveness of the video, they do not provide any means of isolating critical events that elicit strong reactions among the viewers. This latter role is the benefit of the CRM methodology.

4.3 Conclusions

The purpose of this paper was to discuss our initial prototype of a continuous response evaluation system for internet video clips. In summary we found that the prototype was effective in demonstrating the potential for this type of a tool, however despite its face validity, the CRM methodology brings with it all the analytical baggage and difficulties of interpretation that have plagued so many previous investigators. Still, the method offers more potential for in-depth user analysis than the current five-star rating system found on Youtube.com.

The method offers a way of tagging points of interest within video clips, which might be combined with basic demographic information to note for example that "male viewers found segments A, B and C more interesting while female viewers found these other segments X, Y and Z more interesting". This internal metadata could lead the way to selective viewing or "compression on demand" of video segments by information seekers having neither the time nor the interest to view a video in its entirety. This would perhaps be of more value when viewing archival footage of longer video events, such as political speeches, debates, lectures, scientific presentations or videoconferences. The authors also see potential in using the technique in combination with user annotations, class notes and other social indexing artefacts.

References

[1] Alexa Top Sites. Internet:http://www.alexa.com/topsites, [May 19 2009].

[2] G. Richards and N. Ostashewski, “Strategies for success: Meta-cognitive vidcasts for orientation of online learners,” in Proceedings, EDULEARN09, Barcelona, 2009 (in press).

[3] B. Gunther, Media Research Methods: Measuring audiences, reactions and impact. SAGE Publications, 2000.

[4] F. Biocca, D. Prabu, and M. West, “Continuous Response Measurement (CRM): A computerized tool for research on the cognitive processing of communication messages,” in Measuring Psychological Responses to Media Messages. A. Lang, Ed. Lawrence Erlbaum Associates, 1994.

[5] R. Nickerson. “Personal Communication,” Demonstration of the PEAC Video Evaluation System. 1980.

[6] J. Baggeley, “Continual Response Measurement: Design and Validation,” Canadian Journal of Educational Communication, vol 16, no. 3, pp. 217-38, 1987.

[7] J. Maier, M. Maurer, C. Reinemann, and T. Faas, “Reliability and validity of real-time response measurement: A comparison of two studies of a televised debate in Germany,” International Journal of Public Opinion Research, vol 19, no. 1, pp. 53-73, 2006.