A Psychophysiological Approach to the Usability Evaluation of a Multi-view Video Browsing Tool

Carmen Martinez-Peñaranda[1], Werner Bailer[2], Miguel Barreda-Ángeles[1], Wolfgang Weiss[2], Alexandre Pereda-Baños[1]

[1] Barcelona Media, Barcelona, Spain
{maikaranda@gmail.com, miguel.barreda@barcelonamedia.org, alexandre.pereda@barcelonamedia.org}
[2] JOANNEUM RESEARCH – DIGITAL, Graz, Austria
{werner.bailer@joanneum.at, wolfgang.weiss@joanneum.at}

Abstract. The aim of this study is to investigate the usability of a video browsing tool. The tool aims at facilitating content navigation and selection in media post-production, supporting also multi-view content. Psychophysiological measures such as skin conductance level are used to measure cognitive effort. Objective measures based on content retrieval tasks as well as self-report measures of usability are also reported. Results indicate the differential effect of introducing specific support for multi-view content in the browsing tool, and encourage further research on the use of psychophysiological techniques in usability evaluations.

Keywords: psychophysiology, video browsing, usability, interactive search

1 Introduction

With the increasing amount of multimedia data to be handled in production and post-production, there is growing demand for more efficient ways of supporting exploration and navigation of multimedia data. In post-production environments, users typically deal with large amounts of audiovisual material, such as newly shot scenes, archive material and computer generated sequences. A large portion of the material is unedited and often very redundant, e.g. containing several takes of the same scene shot by a number of different cameras. Typically only few metadata annotations are available (e.g. which production, which camera, which date). The goal is to support the user in navigating and organizing these audiovisual collections, so that unusable material can be discarded, yielding a reduced set of material from one scene or location available for selection in the post-production steps. Recently, boosted by 3D cinema and 3DTV, content is increasingly shot with stereo cameras or even more views, recording multiple views of one scene. While this increases the freedom in post-production (e.g., for inserting or moving objects, or adjusting the depth level), multi-view content increases the problem of content management in post-production, as even more (and more redundant) items need to be handled.

Multimedia content abstraction methods such as browsing tools are complementary to search and retrieval approaches, as they allow for exploration of an unknown content set, without the requirement to specify a query in advance. This is relevant in cases where only few metadata are available for the content set, and where the user does not know what to expect in the content set, so that she is not able to formulate a query. In order to enable the user to deal with large sets of content, it has to be presented in a form which facilitates its comprehension and allows judging the relevance of segments of the content set. In order to support users in post-production, media content abstraction methods shall (i) support the user in quickly gaining an overview of a known or unknown content set, (ii) organize content by similarity in terms of any feature or group of features, and (iii) select representative content for subsets of the content set that can be used for visualization. The focus of this work is on assessing the added value of support for multi-view content in order to support users in media post-production.

In this paper we present an approach for evaluating a video browsing tool for multi-view content collections using psychophysiological methods. In Section 2, we give an overview of related work on evaluation approaches. Section 3 discusses the video browsing tool to be evaluated and the task design. We present the results of the experiments in Section 4 and draw conclusions in Section 5.

2 Related work

2.1 Evaluation of Video Browsing and Interactive Video Search Tools

Most of the literature on evaluation of exploratory search deals with text documents. In the multimedia domain evaluation approaches for summarization and skimming systems often deal only with single multimedia documents, rather than with collections. The following classes of evaluation approaches have been proposed (of course combinations of the methods from different classes are sometimes used).

Survey (Self Report). The users are asked about their experience with the tool, their satisfaction with the results and the relevance of certain features of the tool (e.g., [10]). This type of evaluation does not require any ground truth or specific preparation of a data set.

Analysis of System Logs. This approach uses either server-side logs [2] or specific client applications that log user actions [3]. The main advantage is that evaluation does not interfere with the user's work with the system and that the approach can be used for long-term studies. However, comparison across different types of tasks and systems might be difficult.

Question Answering. Users are asked fact finding questions about the content in order to evaluate whether they have found the correct segment of content or were able to extract information from the collection of multimedia documents (e.g., [4]). The questions can be open or in the form of a multiple choice test (quiz). The correctness of answers to open questions needs to be checked by a human, while multiple choice tests can be very efficiently evaluated once the ground truth for a specific data set has been created.

Indirect Evaluation. The user performs a task using the tool or system. Based on the success of this task the effectiveness of the tool can be measured. The task can for example be a content retrieval task [1] or gathering information from a meeting archive browser [7]. Once ground truth for these tasks has been created the answers can be checked automatically.

In this paper, indirect evaluation using content search tasks is performed, in combination with psychophysiological methods and a set of open questions.

2.2 Subjective and Objective Evaluation Methods

As the terminology in the area is often used confusingly, let's clarify that by subjective methods we imply here self-reported qualitative measures, as those obtained by employing interviews, questionnaires or whatever method in which the user is directly asked about some aspect of his experience with a certain content and/or technology. On the other hand, objective evaluation is meant here to refer to the collection of indirect measures of users' experience, that often reveal reactions that are not consciously available to the user (such as variations in performance, motor behavior, emotional reaction and so on) and that therefore cannot be captured by self-report measures. A common occurrence is to refer to the former as observational evaluations and to the latter as experimental evaluations, though in principle, nothing prevents both of these methodologies to work with either objective or subjective data [9].

In self report, the user informs directly about his conscious perceptions, and therefore is "cognitively mediated" [12], and if users cannot report what they cannot perceive consciously, self-report measures are inherently flawed [6]. What is needed is to adopt a dimensional approach to the concept of user experience measurement, where the different aspects that contribute to an engaging experience are analyzed in consonance with their own characteristics; such aspects are as varied as perceptual quality, comfort, emotional reactions, attentional load, etc. The concept of quality of experience is a complex one, and any evaluation activity will need to take into account which of these aspects is a key in the experience delivered to the user. As regards the work discussed here, we consider that the key aspect to measure here is cognitive effort, we combine qualitative methods with indirect psychological measures, namely, performance measures and psychophysiological measures.

2.3 Psychophysiological Evaluation Techniques

Psychophysiological techniques are a perfectly suited methodology for testing cognitive effort in reaction to audiovisual media tasks in an indirect fashion, as this is the branch of psychology that deals with the physiological basis and indicators of psychological processes. Historically, the term has been reserved for the study of the responses of the autonomic nervous system. However, techniques such as EEG (electroencephalography, the recording of electrical cerebral activity) or fMRI (functional magnetic resonance imaging, the measure of changes in cerebral blood flow) also allow observing the activity of the central nervous system. Furthermore, other techniques that cannot be considered as psychophysiological measures (but also let to obtain information from users in a indirect way) such as eye-tracking, have been used on usability research [15]. In any case, for the study of cognitive effort, and certain attentional reactions such us orientation responses, this is still the method of choice for many researchers in the area. Traditionally, registering of autonomic activity is performed over three main measures:

- *Electrodermal activity (EDA)*: reflects changes in the electrical conductance of the skin due to the activity of the sweating glands induced by the sympathetic system, whose activity is related to the degree of emotional activation and it can also be an indicator of cognitive effort.
- *Electromyographyc activity (EMG)*: It measures muscular activity and is often employed to measure facial muscle activity. Depending on the registered muscle and stimuli conditions it can indicate the emotional valence (positive/negative valuation), discomfort and attentional capture.
- *Cardiac and respiratory rhythms*: indicators of phasic attentional responses as well as general stress levels.

These variables are registered by means of the polygraph, an instrument allowing capturing, modulating, amplifying and graphically registering these physiological systems. Regarding cognitive activation measurement, and our research objectives, EDA measures, are often used as cognitive activation and attentional indicators of the level of cognitive effort induced by audiovisual material [11]. Variations in the EDA levels reflect variations in arousal [**Fehler! Verweisquelle konnte nicht gefunden werden.**], which can signal increases in emotional activation or cognitive effort, depending on the stimulation conditions. In our research we focus on exploring the capacity of these technique for measuring cognitive effort.

3 Video Browsing Tool Evaluation

3.1 Tool Description

The developed video browsing tool performs automatic content analysis of newly ingested data. Currently, camera motion estimation, visual activity estimation, extraction of global color features and estimation of object trajectories are performed. In

order to select content, the users follow an iterative selection process, consisting of alternating steps of clustering and selecting subsets of the current data set. Importantly, a new feature was introduced in the latest version enabling to handle multi-view content. This allows users, especially in 3D productions, to reduce the material they have to deal with, and therefore facilitating the search process by reducing the attentional load of the task. The tool does not support text-based queries. The central component of the video browsing tool's user interface is a light table, which shows the current content set and cluster structure using a number of representative key frames for each of the clusters, visualized by colored image borders. The following clustering features are implemented in the tool.

- *Motion activity:* clusters by the global movement in a given clip
- *Color layout:* clusters global color distribution of key frame
- *Media item:* shows all the different video files
- *Multi-view camera:* shows all videos shot with each of the possible cameras
- *Multi-view media Item:* shows all videos, but each view (camera) separately

The browsing workflow starts with selecting a dataset. Then, by selecting one of the available features the content is clustered according to this feature. Depending on the current size of the content set and the available space in the browser window, a fraction of the segments (mostly a few percent or even less) is selected to represent a cluster. The user can then decide to select a subset of clusters that seems to be relevant and discard the others, or repeat clustering on the current content set using another feature. In the first case, the reduced content set is the input to the clustering step in the next iteration. If no key frame of the selected view is available then an alternative view is selected where the user gets informed by displaying the name of the alternative view. On the left side of the application window the history, the similarity search and the result list are arranged. The history feature automatically records all clustering and selection actions done by the user. By clicking on one of the entries in the history, the user can jump back to the selected point. The browsing path can be branched at this point and explored using alternative cluster features without losing previous iterations. To execute the similarity search, the user drags the desired key frame into the marked area for the similarity search. Then, the application displays all available search options to allow the user to further proceed with the search task. The result list is used to memorize video segments and to extract segments of videos for further video editing, for example, as edit decision list (EDL). The user drags relevant key frames into the result list at any time, thus adding the corresponding segment of the content to it. Specific clustering features and further options are offered to the user via a context menu of the key frames where the users can select and discard a whole cluster or select and discard a single video. Furthermore, the metadata of the selected key frame can be displayed.

3.2 Evaluation Measures

The design of the series is based on item retrieval tasks, where the participant has to locate a given item (known or unknown) in a dataset, and performance of the participants is observed under conditions in which the new multi-view feature is used or not. The measures obtained in this tests are the performance measures employed earlier in [1], namely precision, recall, and F1 (the harmonic mean of precision and the recall) of the retrieval task; a psychophysiological measurement of mental activity related responses (EDA, as described earlier). A questionnaire was also filled by every participant in order to gather qualitative reports of their experience with the tool.

3.3 Task Design

For this test we devised four tasks which, we predicted, would vary in difficulty due to the number and the length of the available results. Participants were instructed to use the browsing tool to locate all segments that matched the given textual description, and to note that for two of the four tasks they were not allowed to use the "multi-view" clustering feature. For every task, they had a time limit of ten minutes. The two a priori easy tasks where task A (look for segments with various people walking around) and task C (look for segments showing empty tables with no people present on the scene). The first one had eleven possible solutions with large average segment durations, whereas the second had five solutions but still long segment durations. The tasks where selected also to be more amenable to clustering (by motion activity in the case of task A and by color layout in the case of task C). The two a priori difficult tasks where task B (look for segments showing a man flashing a light in his hand), and task D (look for segments showing two people standing on a platform that falls through an abyss), these had respectively six and one matching segment and the event searched was always of a very short duration. The clustering features were described to the participants and they were introduced to the workflow with the browsing tool. All the participants performed the four tasks, and order of presentation of the task and the presence of multiview condition were counterbalanced between participants. The main condition in the test was the use or not of multi-view clustering, though other variables of interest collected in the questionnaire are the familiarity with the dataset, with web/video browsing general, and with this browsing tool in particular.

4 Results

4.1 Retrieval Performance

We report here the results from the fourteen participants (28 years old average, standard deviation (SD) = 2.7) in the test. Regarding precision and recall measures, overall mean precision was .81 (SD = .33), whereas overall mean recall was .53 (SD = .33). The precision and recall results by task are depicted in Fig. 1. Mean F1 was .61 (SD = .32).

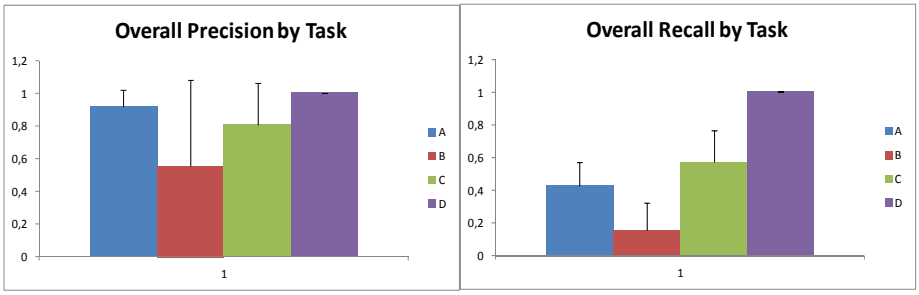

Fig. 1. Precision and recall for the four tasks. Error bars represent standard error

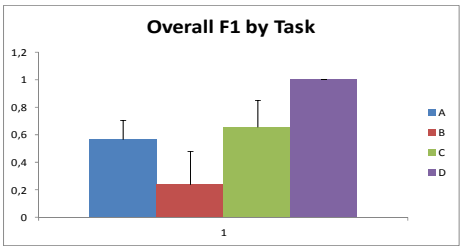

Fig. 2. F1 scores for the four tasks. Error bars represent standard error.

From here on we report only the results of the F1 measure. Fig. 2, depicts the F1 score for each task. As can be seen, the only task that did not match our predictions regarding difficulty was task D, due to the fact that, despite having a single solution, the key frame containing the solution appeared repeatedly represented in various clusters and was easily localizable. Therefore, given that there is no variability associated to that task, we restrict the following analyses to tasks A, B and C. As regards these three tasks, the reasons used to predict their difficulty a priori are supported by the performance data. Of the four grouping variables collected in the questionnaire, F1 scores showed a slight trend to be higher only with high levels of general familiarity with video browsing (see Fig. 3). Regarding the main condition on the test, the use of multi-view clustering (see Fig. 4), there was a clear trend towards better F1 scores when multi-view was allowed for the two harder tasks, but the opposite was observed for the easier task. It could be interpreted that for the easier task, the introduction of a further clustering feature hampers rather than improves performance, though this interpretation must be taken with care, given that the difference in difficulty (as reflected by the F1 score) is not too large between tasks A and C.

Regarding task order, a learning effect was not observed on F1 scores, although for the harder task B, there was a clear difference, given that no participant was able to provide a single correct answer when this task was presented first. Evidently, as participants work with the tool, they grow increasingly familiarized with the materials increasing the chances of encountering the relevant segments.

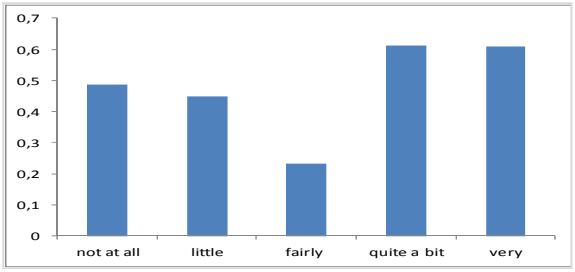

Fig. 3. F1 Scores by familiarity with video browsing in general.

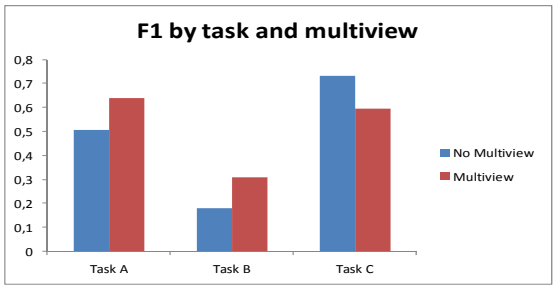

Fig. 4. F1 by task and multi-view.

4.2 Electrodermal Activity Data

EDA measures were also registered in order to obtain a further indicator of the amount of mental effort associated with the browsing activity. As described in above, the increase in arousal signaled by variations in EDA levels has been used as an indicator of mental effort when the emotional significance of the experimental material is controlled as proposed initially by [5]. Thus, we calculated the mean EDA activity for the different tasks and both levels of the multi-view condition. Z-scores (standardized scores) of EDA values for each participant in each task were calculated from the raw data by subtracting the mean EDA value of all tasks to the value of each task, and dividing the results by the standard deviation of all task of this participant. Regarding the different tasks there were no significant differences between the levels of EDA for the different tasks (F<1, see Fig. 5a), nor for the order of the tasks, that is, during the approximate one hour session, the levels did not vary with time (F<1, see Fig. 5b).

Regarding the multi-view condition though, a clear trend was observed for a higher activity in conditions where multi-view was allowed (F = 2.64, p < .078, see Fig. 6a), which was mostly due to differences between tasks B and C (respectively F = 2.8, p < .20, and F = 2.72, p < .118, see Fig. 6b). The fact that the comparisons fell short of significance is clearly due to the few participants tested and a few participants more will surely provide significant results.

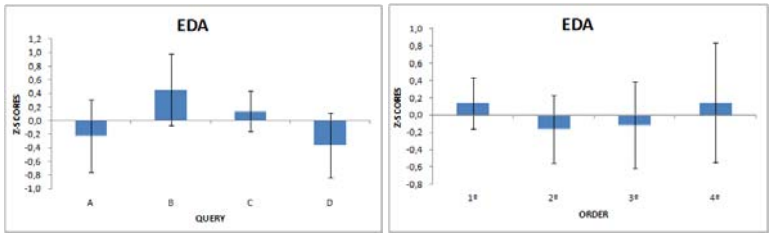

Fig. 5. (a) EDA values by task (query), (b) EDA values by order of presentation. Error bars represent standard error.

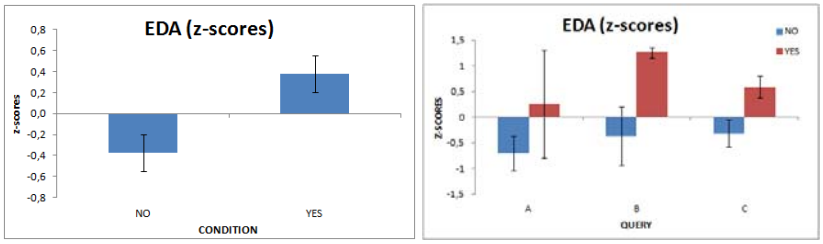

Fig. 6. (a) EDA average values per multi-view condition, (b) EDA average values per multi-view condition and task Error bars represent standard error.

It is evident that the EDA data further corroborates the results from the performance test, and this with just an average analysis of tonic activity. These results are a further proof of the impact that indirect measures of user's reaction might have for usability tests. The fact that the EDA activity is higher for the condition where multi-view was allowed might seem counterintuitive at first, if we think that a better performance should be related to less mental effort. But the concept of mental effort is not necessarily negative when performing a task. More complex tasks require more mental resources (i.e. attention, or cognitive effort) to be processed [5], and it can be reflected on an increased arousal. It does not mean that the mental processing of the more complex task has to be hindered respect to a simpler one. For example, Lang et al. [13] found that increasing the structural complexity of television messages (by increasing the number of edits on them) resulted on increased arousal on viewers, more attention paid (measured by a decrease on heart rate) and even a better memory for the content of the messages. These authors explained that more structurally complex stimuli elicit automatic orientation of attention [14], which, in turn, increases the automatic allocation of mental resources to the task processing, and thus improving it. However, if the complexity of the stimulus is excessive, the mental resources required to its processing exceed the mental capacity of the viewer, resulting in a poor processing and a bad recall of the content. Some similar could be happened at our experiment: The inclusion of the multiview condition (i.e. the increase on the complexity of the task) might have required more attention (i.e. "cognitive effort") to the correct accomplishment of the task, automatically mobilizing more mental resources than the simpler condition, and so providing better F1 results, while the more complex tasks

(e.g. task B) also demanded more attentional resources, but overloaded the cognitive resources availability of users, resulting in poor F1 results. In any case, these results are only preliminary evidence and further research is needed to assess how the concepts of cognitive effort and physiological arousal can be employed in usability evaluations. But the fact that we have observed this phenomenon opens interesting avenues for future research on usability and browsing behavior.

4.3 Verbal Reports

Participants suggested a series of features for improving the usability of the tool, for example, customizable window size, showing frames on the timeline of the video player or the possibility to select various videos at once. Participants did not like that in occasions the clusters are automatically modified when changing frame size, some did not like the clustering features available, or the overlapping between the preview and video player windows. As it is well known from evaluations of previous versions of the tool, the fact that all participants find unanimously distracting is that the key frames for a given clip are not constant, but depends on the feature used for clustering. We found that we had to emphasize the use of discard and select clusters function in the context menu. A participant suggested making these actions more evident by putting them on the above menu bar. Turning finally to the positive feedback, those who were not familiar with this kind of search liked the idea of searching videos by visual features and, as one participant put it; its potential utility in editing works where the content of the clips is not directly relevant and search by primitive visual features suffices. Also, they all unanimously agreed that they were more comfortable with using the similarity search than clustering.

5 Conclusion

The first thing to note is that participants tended not to produce false alarms, as reflected by the asymmetry between high precision and low recall. The explanation probably lies in the moderate size of the dataset. The F1 score behaved as expected in terms of the predicted difficulty of the queries, and the main result to emerge from this study is of course the differential effect of introducing the multi-view condition, which was further confirmed by the EDA data, although more detailed analyses could shed more light on the exact causes of these effects. For example, looking at phasic EDA responses, which would need an event-related approach to the design of the task, was not possible with the present material. It is interesting to note how the EDA measure was sensible to the manipulation of the search processed introduced by allowing the use of multi-view, but that this was not the case for the manipulations of task difficulty in terms of number and size of the relevant segment. This could be interpreted as a further confirmation that the variations in EDA level observed are due to variations in the attentional load of the task, rather than an increased emotional activation due to the difficulty in finding relevant results, though again, an empirical confirmation of the latter would require a more controlled task design. Finally, re-

garding the experience of the participants with this kind of software, only the previous experience with video browsing in general seemed to have an effect on performance, though a bigger pool of participants would be needed to extract firm statistical conclusions in this regard. It is also very interesting to note that the verbal feedback provided by the participants after performing the task made few direct references to the effect of having or not the possibility of using multi-view clustering, despite the fact that EDA and precision/recall measures indicate that this is a key variable affecting their performance. This is an encouraging result in terms of bringing this kind of measurement into evaluating tools professionals' use in their daily work.

Acknowledgements. The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 215475, "2020 3D Media – Spatial Sound and Vision".

References

1. Werner Bailer and Herwig Rehatschek. Comparing fact finding tasks and user survey for evaluating a video browsing tool. In Proc. of ACM Multimedia, Beijing, CN, Oct. 2009.
2. S. Fissaha Adafre and M. de Rijke. Exploratory search in Wikipedia. In SIGIR Workshop on Evaluating Exploratory Search Systems, 2006.
3. B. J. Jansen, R. Ramadoss, M. Zhang, and N. Zang. Wrapper: An application for evaluating exploratory searching outside of the lab. In SIGIR EESS Workshop, 2006.
4. V.B. Jijkoun and M. de Rijke. A pilot for evaluating exploratory question answering. In SIGIR Workshop on Evaluating Exploratory Search Systems, 2006.
5. Kahneman, D. *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall, 1973.
6. Knoche, H., De Meer, H. G., y Kirsh, D. "Utility curves: Mean Opinion Scores considered biased". Proc. of 7th Intl.Workshop on Quality of Service. London, June 1999.
7. W. Kraaij and W. Post. Task based evaluation of exploratory search systems. In SIGIR Workshop on Evaluating Exploratory Search Systems, 2006.
8. Ravaja. N. Contributions of psychophysiology to media research: Review and recommendations. *Media Psychology, 6*, 193-235, 2004.
9. Jumisko-Pyykkö, S., Strohmeier, D. "Report on research methodologies for the experiments", Technical report of the MOBILE3DTV project, 2008.
10. Yan Qu and George W. Furnas. Model-driven formative evaluation of exploratory search: A study under a sensemaking framework. Inf. Process. Manage., 44(2):534-555, 2008.
11. Sánchez-Vives, M. V., and Slater, M.. From presence to consciousness through virtual reality. Nature Reviews Neuroscience, 6, 333-339, 2005.
12. Wilson, G. M., & Sasse, M. A. "Listen to your heart rate: counting the cost of media quality". Affective interactions towards a new generation of computer interfaces. 2000.
13. Lang, A., Zhou, S., Schwartz, N., Bolls, P. D., & Potter, R. F.. The effects of edits on arousal, attention, and memory for television messages: When an edit is an edit can an edit be too much? *Journal of Broadcasting & Electronic Media, 44*, 94-109, 2000.
14. Turpin, G. Effects of stimulus intensity on automatic responding: The problem of differentiating orienting and defense reflexes. *Psychophysiology, 23*, 1-14, 1986.
15. Cooke, L. Eye tracking: How it works and how it relates to usability. Technical Communication, 52, 456-463, 2005.