Techniques for Evaluating Novice-Oriented Creativity Support Tools

Nicholas Davis,
Alexander Zook,
Mark Riedl
School of Interactive Computing
Georgia Institute of Technology
{ndavis35; a.zook;
riedl}@gatech.edu

Friedrich Kirschner
Hochschule fuer Schauspielkunst
Ernst-Busch
Berlin, Germany
f.kirschner@hfs-berlin.de

Michael Nitsche
School of Literature, Media, and
Communication
Georgia Institute of Technology
michael.nitsche@gatech.edu

ABSTRACT

We present evaluation techniques for a novice-oriented creativity support tool in the domain of digital filmmaking. Novices need help *executing* tasks as well as *knowing* which tasks are appropriate and the implications their decisions have for their creative product. With our tool, we focus on supporting critical domain knowledge that can enable novices to make meaningful creative contributions. In film, domain knowledge includes cinematographic and editing rules and conventions. Our creativity support tool (CST) provides feedback when novices violate these norms. We use two approaches to evaluate the results: (1) Expert Consensual Assessment, and (2) Individual-Group Consensus, which is a specialized technique we developed to overcome the limitations of the Consensual Assessment Technique for this application. We discuss these two techniques and their domains of application. Finally, we call for further research into evaluation techniques for CSTs sensitive to the task evaluated (e.g. execution vs. knowledge support) and relevant domain (e.g. machinima).

Author Keywords

Creativity Support Tools; Evaluation; Cognitive Science

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms

Human Factors; Design; Measurement.

INTRODUCTION

In this paper we investigate evaluation techniques for tools to support creative digital filmmaking. We study machinima, which is a new form of digital filmmaking that leverages the real time graphic rendering capabilities of video game engines to create polished animations. Machinima films began as a recording of scripted video game characters with audio overlay. However, the tools of the trade have expanded beyond the initial confines of early video game engines to introduce much more complexity and nuance and include control of lighting, set and character design, and cinematography. This technology adaptation has opened a door to individuals with no experience in animation or filmmaking to create professional looking animated films. The proliferation and open nature of machinima tools have introduced many new avenues for creative expression and significantly lowered the barrier to entry for digital filmmaking. These tools have empowered individuals to creatively express themselves in ways that were prohibitively expensive a decade ago. However, despite how powerful the tools have become, many elements of the filmmaking craft remain unknown to novices.

There is a filmic language that denotes standards, conventions, and general cinematographic rules that experts gradually learn through education and experience [1]. Although experts often violate these rules for stylistic purposes, novices unknowingly violate these rules, which can interrupt the visual continuity between shots, temporal rhythm, and spatial orientation of a scene.

Our earlier study [3] found novices routinely violate many of these filmmaking conventions, and the creativity of their products suffers. Although the graphics of their film may look sophisticated, creative decisions that unintentionally violate cinematography and editing conventions distract the viewer from the story. The study reveals novices require creativity support in two ways: (1) *executing* known creative tasks and (2) *knowing* the norms and values of a creative domain. How can we help novices avoid errors associated with execution and knowledge to produce higher quality creative content?

Lubart [9] enumerates four ways computers can support creativity. A computer *nanny* provides tools to schedule and maintain creative activities. A computer *pen-pal* supports collaboration within teams. Computer *coaches* stimulate creative thinking by suggesting creative activity based on expert knowledge in the domain. Computer *colleagues* may meaningfully contribute to tasks so a human-computer team becomes a contributor to a domain.

Novices require creativity support tools (CSTs) that provide guidance like a coach and also perform skilled operations like an expert colleague that simplifies the process of creation. Execution of a known creative goal is not sufficient support for novices. Rather, coaching must support novices acquiring the deep domain knowledge experts possess. Lacking this knowledge, novices are unable to understand the implications of their decisions for achieving creative outcomes. Novices are often unable to contextualize their creative goals within a new domain without this knowledge.

The video game engine in the machinima tool performs the expert task of animating content, but users also need a coach that provides feedback about their decisions. To address this need we designed an intelligent creativity support system that analyzes camera placement to determine if the user has violated cinematographic rules, such as those listed below.

- **180 Degree Rule:** The line of action is an imaginary line created between two individuals engaging in a conversation. Once a camera is placed on one side of the line of action, subsequent camera angles should remain on that side of the line of action to prevent the characters from changing their perceived location [7].
- **30 Degree Rule:** The degree of change between two sequential camera angles should be greater than 30 degrees to reduce jumpy or jittery cinematography [5].
- **Cutting on Action**: Camera angles should switch during an action being performed, rather than immediately before or after the action. This creates the illusion of fluid movement [2].
- **Pacing**: Shots should change at a regular frequency and abrupt changes should only occur during highly emotional or dramatic moments [8].

Unlike a simple error-correction approach our tool aims to help inform novice decisions. Feedback provides information and explanations of violations to inform users without constraining them to follow the norms of cinematic practice. Existing machinima tools support novice *execution* needs; our approach targets novice *knowledge* needs. We balance the creative freedom of the user with the need for domain-specific knowledge.

In the remainder of the paper we first describe an experiment to evaluate our machinima CST. We then present two techniques to evaluate the novice films created in our study to understand the impact of our CST. This analysis addresses successes and limitations of the approaches. We conclude by discussing the broader application of one of the techniques and call for further analysis of CST evaluation techniques specific to evaluation tasks – such as execution or knowledge support – and evaluation domains – such as machinima creation.

Experiment Design

We will only briefly outline the experiment design here since this paper focuses on CST evaluation techniques. Our experiment investigated the effect of offloading expert knowledge about rules onto digital filmmaking CSTs [4]. We hypothesized that providing knowledge about the filmic language would help users make more informed decisions as they relate to rules, manifested in fewer violations of those rules. Our study found this to be the case, see [3] for more detailed results.

Evaluating systems that support novices is difficult because they lack baseline execution skill and knowledge in the application domain. Novices may have difficulty executing a desired task using the interface and knowing appropriate domain content to use. To minimize execution problems, we isolated and simplified the required functions for the task. Participants were only responsible for selecting the specific angle and timing for a shot from a list of predetermined camera angles. This limited the learning curve for approaching the software and focused our evaluation on novice cinematic knowledge support.

We recruited 20 participants for this study. Participants were split into two groups that both engaged in a constrained creative activity using the Xtranormal machinima software (www.xtranomral.com). We selected Xtranormal because it is a popular and freely available machinima creation tool. A pre-scripted scene and predefined set of expert selected camera angles were provided to the participants to reduce the amount of training and tool expertise required for the task. The task was to find the best editing and cinematography for the scene with the given materials, which meant selecting the appropriate camera angle and deciding when it should be used in the scene. Users could insert or remove any number of camera angles from the list during the 40 minutes they were provided for the task. The control group (n=10) was unaided. The experimental group (n=10) was able to press an ‘Analyze’ button that prompted the system to analyze their camera selections and provided feedback about any rules they violated.

The program that analyzed the films was a Wizard of Oz (WOZ) system in which a human expert evaluated the user’s decisions and sent standardized feedback to the user based on the rules that they violated. Users were able to request feedback on their current selections from the system whenever they desired. Feedback was provided through an IRC chat channel that appeared next to the Xtranomral interface. The participants thought the feedback came from an automated system.

The ‘wizard’ watched each film in real time and noted all current errors. The wizard’s assistant entered these errors into an interface using template text explanations. When participants pressed the ‘Analyze’ button on the WOZ interface, the assistant sent a message that contained a list of the user’s rule violations as well as explanations describing those rules, similar to the explanations above.

Evaluation Techniques

We evaluated our CST by comparing the resulting novice films using both a standard expert evaluation methodology and a specialized method developed for our domain. Evaluation focused on novice errors as our previous study found these are required to recognize creative success in the machinima domain [3]. Below we describe the two consensus-based evaluation methods we used on our CST and their strengths and limitations.

Expert Consensual Assessment

One approach to evaluate the effectiveness of our CST is to use the Consensual Assessment Technique (CAT) [6] where a panel of experts rates the users’ products. We attempted this evaluation by recruiting three film experts to independently rate the general and technical quality of each of the film clips. However, these rating had a low inter-rater reliability and were therefore discarded as invalid. There are several factors that may have contributed to this result.

One explanation could be that CAT may be less appropriate when the creative products are extremely similar. Since we studied novices who were not familiar with the machinima tool or creating films, we simplified their contribution in way that minimized the necessary skill. This also changed the degree of creative freedom that participants had to create diverse products. The experiment was designed to constrain the creative task to selecting the angle and timing of camera placement from a pre-defined list of camera angles. The set of film clips created were approximately 1-minute long and appeared very similar to each other visually. The camera angles had significant variations, but the overall visual similarity of the scenes may have reduced the evaluators’ sensitivity. It may have been hard to make meaningful comparisons between the clips.

A second explanation could be that evaluating film is fundamentally different than evaluating static images or products (as is more traditional in CAT) because they cannot be placed next to each other for comparison. Timing effects may come into play since the product is experienced through time. To mitigate this, we encouraged evaluators to take notes of important information. However, sequential evaluations can skew expert ratings due to anchoring on early aspects of a film clip or earlier clips in a series [10].

A third factor is that the time and focus required for a series of ratings may lead to evaluator fatigue. Evaluating films typically takes more time than evaluating static images as each film has to be watched in its entirety. More rapid fatiguing makes it difficult to reach consistent evaluations from a set of experts when examining more than a few movie clips.

Individual-Group Consensus Evaluation

To bypass the limitations of CAT in our domain we developed a method of gathering individual evaluations followed by a group consensus process. Three researchers who helped design the experiment analyzed each film to determine rule violations. It was important to select individuals familiar with the design because they would be able to effectively detect rule violations. The analysts were familiar with the pre-determined camera list and had a sense of which shot combinations typically constitute an error.

Each analyst watched all the clips (in random order) independently and noted any errors. There were four rules in total, and each shot could have multiple errors. The analysts watched each individual shot and noted the errors they detected individually. Using analysts familiar with the domain, aware of experimental constraints, and taking detailed notes on error timing helped circumvent the limitations of applying CAT discussed above.

The analysts aggregated their error judgments for a combined evaluation. Multiple possibilities for combining these ratings exist. One option is to use a majority rules voting system for the error data for each shot. In a majority rules approach, whichever decision had at least two-thirds support would be selected. This approach is faster because it is accomplished without viewing the film clips again and occurs without much discussion. However, it runs the risk of overlooking crucial data due to a shallow analysis of differences in interpretation among raters. When employing CAT we found experts would agree on qualitative classes of errors, but differ in the specific points they found problematic. A majority rules approach suffers this same limitation.

We decided to instead have all three analysts view all the clips in their entirety a second time as a group. Each clip was analyzed fresh and the analysts came to a consensus agreement on each clip. The group decision was then compared to each of the individual analysts’ error evaluation. If there was a discrepancy between the group decision and any of the individual analysts’ error evaluation, a further investigation was pursued. The analyst(s) whose error evaluation contradicted the group decision was asked to explain why he or she made that decision, and if s/he still felt the same way. The group would then discuss this information and decide which decision to maintain.

We preferred this approach because detecting errors was difficult given that each 1-minute scene contained between 5-15 shots and there were approximately 30 clips. It is easy for one person to overlook an error. Introducing multiple viewers reduced the likelihood of missing an error. Additionally, the individual data helped to highlight potentially inaccurate decisions the group came to. There were more redundancies in this approach, which is preferable because it presented more situations to challenge and rectify the evaluation decisions. We found the group consensus approach helped detect many overlooked errors and correct for differing interpretations among evaluators, particularly in regards to pacing rule violations.

CONCLUSIONS

Novices often lack both the skills to *execute* a desired creative goal and the *knowledge* to be aware of whether their execution is within the norms and values of a domain. Evaluating creativity support tools that are targeted for supporting knowledge and reducing fundamental errors may require specialized evaluation procedures. Our CST was designed to support the informed exploration of camera angles and provide feedback to the user about when their decisions violated important cinematic norms. This type of feedback encodes expert knowledge into the CST and enables novices to achieve better creative outcomes. We argued that testing the effectiveness of this type of support is best measured using an individual-group consensus approach with detailed product analysis.

Several factors may have contributed to invalid CAT judgments, including the similarity of products, the temporal nature of film, and evaluator fatigue. We circumvented these limitations through a modified technique that retained expert evaluation and consensus while introducing greater awareness of task constraints, a round of detailed and recorded evaluation, and a consensus approach focused on aligning group and individual assessments. The individual-group consensus process adds much needed redundancy to creativity evaluation with more careful accounting of disagreements.

The technique we describe is applicable to a wide variety of creative tasks, but is particularly useful when examining temporally extended or “large-scale” artifacts. CST evaluation often requires detailed information on the effects of a CST intervention on a given product, rather than only a holistic judgment of the resulting changes in user creativity. In our case detailed judgments of knowledge-related errors were required; other domains likely require other metrics. An individual-group consensus approach centered on these detailed evaluation points can bypass cognitive biases (e.g. anchoring) and limitations (e.g. fatigue) that hinder the use of holistic judgment in these domains. A further benefit of this method is providing detailed information to inform the further development of CSTs, documenting the relative magnitude of different aspects of creativity in a target domain. Ultimately, we hope this work highlights the need to develop CST evaluation techniques that are sensitive to the evaluated task (execution vs. knowledge support) and evaluation domain (e.g. machinima).

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1002768. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

1. Arijon, D. (1976). Grammar of the film language.
2. Beller, H. *Handbuch der Filmmontage. Praxis und Prinzipien des Filmschnitts.* Tr Verlagsunion , München, 1999.
3. Davis, N., Zook, A., O’Neill, B.O., Grosz, A., Headrick, B., Nitsche, M., Riedl, M. Creativity Support for Novice Digital Filmmaking. To Appear in *Proceedings of Human Computer Interaction (CHI) 2013.*
4. Davis, N., Li, B., O’Neill, B., Nitsche, M., and Riedl, M. Distributed Creative Cognition In Digital Filmmaking. In *Proc. Creativity & Cognition 2011*, ACM Press (2011), 207–216.
5. Hayward, S. *Cinema Studies: The Key Concepts*. Routledge, London, 2000
6. Hennessey, B. A., Amabile, T. M., & Mueller, J. S. (1999). Consensual assessment. *Encyclopedia of creativity*, *1*, 346–359.
7. Katz, S. D. *Film Directing Cinematic Motion: A Workshop for Staging Scenes*. Michael Wiese Productions, Studio City, 2004.
8. Kindem, K. and Musburger, R. *Introduction to Media Production: The Path to Digital Media Production.* Focal Press, Burlington, 2009.
9. Lubart, T. How can computers be partners in the creative process: Classification and commentary on the Special Issue. *International Journal of Human-Computer Studies* 63, 4-5 (2005), 365–369.
10. Tversky, A. and Kahneman, D. (1974). "Judgment under uncertainty: Heuristics and biases". *Science*, 185, 1124–1130.