How Do We Evaluate the Quality of Computational Editing Systems?

Christophe Lino, Rémi Ronfard, Quentin Galvane, Michael Gleicher

How Do We Evaluate the Quality of Computational Editing Systems?

Christophe Lino[1], **Rémi Ronfard**[1], **Quentin Galvane**[1], **Michael Gleicher**[1,2]
[1] Inria, Univ. Grenoble Alpes & CNRS (LJK), Grenoble, France
[2] Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI, USA

Abstract

One problem common to all researchers in the field of virtual cinematography and editing is to be able to assess the quality of the output of their systems. There is a pressing requirement for appropriate evaluations of proposed models and techniques. Indeed, though papers are often accompanied with example videos, showing subjective results and occasionally providing qualitative comparisons with other methods or with human-created movies, they generally lack an extensive evaluation. The goal of this paper is to survey evaluation methodologies that have been used in the past and to review a range of other interesting methodologies as well as a number of questions related to how we could better evaluate and compare future systems.

Introduction

Automatic film editing has a long history, dating back at least to Gilles Bloch's PhD thesis in 1986 (Bloch 1986). However, evaluating editing systems remains an open problem which makes it difficult to measure progress of the field as a whole. In this paper, we focus on the question of how the community has evaluated editing systems in the past, what their limitations are and what alternative methods of evaluation could be proposed to remedy those limitations

This paper is organized as follows. We first look at the reasons why our community should be interested in evaluating film editing. Then we explain what makes evaluation difficult. We then review methodologies which could be useful in the future. This is given as a list of alternatives. Should we use objective or subjective evaluations ? What should we measure? How can we design valid empirical studies? After reviewing some possible answers, we conclude by stressing the importance of sharing data sets and codes.

Why evaluate ?

From a methodological perspective, we feel it is important to understand why we need to evaluate our work and what we can expect from such evaluations, since this may have an impact on which methodology is most appropriate to reach those goals (Gleicher 2012).

Evaluation is an important part of a research agenda to understand whether the goals are being met. Thorough consideration of evaluation can serve to help clarify these goals, and to encourage making them explicit. Therefore, asking questions of evaluation can also be an aid in directing future research towards better-defined goals and a more collaborative research effort to reach them. Evaluation can expose limitations, or successes, of current approaches, also helping to steer progress.

When evaluation can be made to support comparison, especially when it can quantify different results, it can serve a valuable role in measuring progress. For example, it can be useful for comparing methods, understanding tradeoffs (is a more complex method "worth it"?), and assessing respective advantages and shortcomings.

For the automated filmmaking community, the lack of clear metrics for evaluation has been problematic. We have no ways of comparing ourselves to previous work. A primary reason for evaluating automatic editing tools is therefore to compare methods and to measure progress. As we will emphasize later, this calls for shared ressources such as data sets and codes, which have never been available to our community. Even different versions of the same method can be hard to compare. For a human viewer, objectively evaluating the quality of an edit is a difficult task. Comparing two edits of the same story shown side-by-side is difficult since it requires the viewer to watch both at the same time. And comparing the two edits shown one after the other, may introduce a bias through the order in which they are shown.

Evaluation is potentially multi-faceted. In this paper, we focus on the evaluation of "result quality" because it has unique challenges for automatic filmmaking. However, we note that other aspects of systems, such as robustness, resource usage (*i.e.* speed), usability, and ease of implementation are also important concerns in automatic filmmaking.

Why is evaluation hard ?

The difficulty of assessing the quality of film editing can be traced back to at least three reasons. The first reason is that there is never a single correct answer to an editing problem. Starting with the exact same set of rushes and the same set of objectives, different edits may be considered equally correct solutions. As a result, we cannot compare to ground truth. Even trivial solutions such as an extended shot with no

cut can never be dismissed. Problems in film editing usually have multiple solutions, which are all equally valid, though such solutions may be very different from each other.

A second reason is that the quality of editing can never be judged directly. Editing is an invisible art – its effect is indirect. As noted by many professional editors, the best cut often remains unnoticed and invisible (O'Steen 2009). How can we evaluate something which is (in its own account) invisible?

A third reason why evaluating film editing is difficult is that the rules of good editing are not absolute – they can be used as guidelines, but they are neither formally defined nor mandatory. Expert editors break editing rules very often (*e.g.* to introduce an emotional impact on the audience) and, as in many artistic endeavors, the first and foremost rule is actually to communicate to the audience. Current computational systems are good at implementing formal rules, and can be evaluated in this respect. But they cannot easily recognize situations where it is valuable to break those rules.

We should note that those difficulties are not specific to computational systems. In fact, the difficulty of evaluating film editing has also been recognized by professional film editors. An article in The New York Times (Harris 2008) reminds us that

> the invisible art, as many of its practitioners call (film editing), has been an Oscar category since 1934. Yet editors acknowledge that even after 70 years assessing excellence in their field sometimes comes down to guesswork. Everything else - music, cinematography, costumes, design, acting - can be judged at face value. But when you're looking at editing, you don't know what the totality of the material was, and you don't know the working dynamic between a director and an editor. It's very difficult. (...) The answer, editors themselves say, is to avoid the question. When considering a movie for awards, they focus on narrative, pace and (actor) performance.

The last part of this quote hints at the important idea that a better way of evaluating film editing may be to measure how much it contributes to other, more measurable, aspects of film-making such as advancing the story, establishing the pace or enhancing the actor performances. We will get back to this later.

Subjective vs. objective evaluation

Given the difficulties already mentioned, an objective evaluation of film editing does not appear to be possible. One method that we would especially like to rule out first is the use of a ground-truth, *i.e.* comparing automatically edited movies with the work of a human expert who makes no error (the ground truth). In the context of film editing the definition of an "error" is unclear; there are many ways of editing a given scene, and all of them can be correct. In that sense, there is no "ground truth" to compare with.

Another way to provide a numerical value of the quality of an edit could be to proceed with a quantitative analysis of how many editing errors have been made. Psychologists d'Ydewalle and Vanderbeeken offer a useful classification of editing errors (D'Ydewalle and Vanderbeeken 1990). Editing errors of the "first order" are small displacements of the camera or image size, disturbing the perception of apparent movement and leading to the impression of jumping (also called "jump cuts"). Editing errors of the "second order" are violations of the spatial-cognitive representation of the 3D scene (*i.e.* breaking the continuity). One example is the 180° rule violation, where the camera crosses the line between two actors and as a result the actors appear to swap positions. Another example is the motion continuity violation, when the camera crosses the line of an actor's movement and as a result the actor appears to change directions. Editing errors of the "third-order" are when successive shots have too little in common to be integrated into a single chronological sequence of events (*i.e.* non-motivated shots and cuts).

On one hand, such an evaluation can provide a good means to assess the grammatical correctness of an edited movie. On the other hand, it can be argued that jump cuts or violations of continuity rules are sometimes made intentionally by expert cinematographers; it is a mean to create an effect on viewers (Smith 2005), such as introducing tension or making viewers aware that something dramatic is happening. Furthermore, editing errors of all kinds are not equally important, and cannot be compared to each other easily. Thus even the counting of errors does not really provide an objective measure of the quality of the edit.

Film editor Walter Murch enumerates criteria that he uses to evaluate his own work (Murch 1986). They go beyond the presence or absence of editing errors. According to him, the perfect cut must fulfill six criteria which are, in order of importance, emotion (how will this cut affect the audience emotionally at this particular moment in the film?), story (does the edit move the story forward in a meaningful way?), rhythm (is the cut at a point that makes rhythmic sense?), eye trace (how does the cut affect the location and movement of the audience's focus in that particular film?), two dimensional place of screen (is the axis followed properly?) and three-dimensional space (is the cut true to established physical and spacial relationships?). While Murch gives numerical percentages to the six rules, his formula can hardly be used to evaluate a cut objectively, especially since it relies so heavily on emotion and story.

As a result, subjective evaluation appears to be better suited to the evaluation of film editing. Subjective evaluations can be performed by experts or non-experts. To assess or improve an editing system, getting formal feedback from expert filmmakers on the output of a system appears to be of greater value. The use of "golden eyes" has proven useful in other contexts, such as the subjective evaluation of video compression standards, including MPEG and JPEG-2000.

This kind of evaluation however also raises questions. First, experts come with different degrees of expertise. How do we compare their evaluations? Experts may also not agree with each other. Thirdly, experts are more apt to compare automatically edited movies with professionally edited movies, rather than different automatic methods (which they tend to dismiss equally). Thus the comparison between professional editing (represented by experts) and automatic editing can turn out to be counter-productive. One final as-

pect to account for is that professional editors may also be unsupportive of the notion of automatic editing altogether.

Empirical Study of the Effects of Editing

In the previous sections, we considered the challenges of assessing editing directly. As we discussed earlier, editing is an "invisible art" (O'Steen 2009; Harris 2008; Apter 2014) that is often not consciously noticed by the viewer. However, that doesn't mean that it does not affect the viewer. If we did not expect to have some effect on the viewer, there would be little reason to do the work. Editing, as part of the filmmaking process, is done to achieve some goal in terms of the results on the viewer. By directly measuring these results on the viewer, we can indirectly evaluate the success of editing. Our distinction here is between goals that are directly about the filmmaking techniques themselves (*e.g.* trying to mimic the editing style of a famous editor) and goals that indirectly assess the techniques by measuring the success of the film itself (*e.g.* establishing pacing or conveying information).

Such indirect evaluation of editing, by measuring its effects rather than directly observing its qualities, has two inter-related issues. First, we must identify what effects to measure. Second, we must design experiments that measure these effects in ways that allow us to attribute the causality of the difference to the particular aspect of the process that we are interested in (*e.g.* that a measured difference is caused by differences in editing, rather than differences in lighting or the mood of the viewer). Such an indirect approach to assessing editing has advantages including that it considers the ultimate consumers of the results, and relies on human subjects empirical studies for which there is significant experience and methodological development in the social sciences and statistics. Increasingly, such studies are part of computer science, especially in the area of Human Computer Interaction. Here, we focus on some of the unique challenges of performing such studies for evaluation of automatic filmmaking.

While there may be specific effects that editing is trying to achieve, such as creating a sense of rhythm or assisting in helping create clear event boundaries, editing is typically part of the more holistic aspect of filmmaking, and its goals are therefore to assist in achieving the overall goals of the resulting film. Therefore, the goals of editing, are ultimately the goals of the film itself. Some of these goals may be "lower level", such as does the film guide the viewers attention to the place where the director wants it? Or conversely, does the film avoid distracting the viewer from the important aspects? Higher level goals include successfully telling a story, conveying information, or imparting a feeling. These effects may be subjective (*e.g.* did the viewer like the story?) or objective (*e.g.* how much of the presented information did the viewer remember?).

There are many different potential goals for a film, and several have already been used as methods for assessing aspects of the the filmmaking process. For example, in an educational video, one may care about the viewer's recall of the information or their feelings of association with the presenter. One example using these goals for evaluation is the work of Andrist *et al.* (Andrist et al. 2012) that showed how subtle differences in "acting" (how an animated presenter moved its eyes) created measurably different outcomes in how much information the viewers remembered and their assessment of the presenter. Similarly, Ponto *et al.* (Ponto, Kohlmann, and Gleicher 2012) considered how different camera movements flying through a scene affected viewer's recall of objects in a scene, as well as the ease with which the viewers felt they could identify objects. The ability of films to correctly convey event structure has been explored by the perception community (*e.g.* (Magliano and Zacks 2011)), and their experiments may provide a mechanism for comparisons between video production techniques. A more specialized example of a measurable goal is for a horror film to be scary. This two can serve as an evaluation criteria, for example Branje *et al.* (Branje et al. 2014) evaluate the impact of adding tactile stimulation to the "shockingness" of horror movie clips. Specially designed video clips are often used in research to induce moods in experimental participants (Kučera and Haviger 2012) to understand the effects of mood on tasks. These experiments could be turned around to measure the effectiveness of different clips are creating the emotions by using these previously reported task effects.

Jhala and Young (Jhala and Young 2009) propose a methodology to evaluate the effectiveness of a camera system to convey a story. They particularly evaluated the efficiency of their editing system Darshak in communicating the underlying narrative content. Their approach is based on an established cognitive model of story comprehension, which uses a model of stories called QUEST (Graesser, Lang, and Roberts 1991), in which stories are represented as conceptual graph structures, and a psychological model of question-answering which supports questions of types why, how, when, enablement, and consequence. QUEST had been designed to evaluate plan-based computational models of narrative (Christian and Young 2004). Jhala and Young used it to compare different visualization strategies in communicating the story. This appears to be a promising approach for evaluating the effectiveness of film editing in conveying stories.

The identification of an effect to measure is different than the choice of a method to measure it. For example, consider the goal of imparting a feeling of sadness. If we have a video whose intent is to impart a feeling of sadness on the viewer, we can measure its success by measuring the sadness of viewers after watching the film. Such measurement can either be made subjectively (*e.g.* asking them to rate their sadness), or objectively (*e.g.* seeing how their behavior changes based on their mood). However, the two choices are often tightly inter-twined: some goals may be more or less hard to measure.

There are an array of potential mechanisms for performing empirical studies. The primary one is to ask questions of the viewers. Such questioning can either be subjective (*e.g.* rate how much they liked it, or how easy it was to understand) or objective (*e.g.* ask them to perform a task that requires them to have watched the video, such as a quiz on its contents). However, other forms of instrumentation are possible. For example, eye tracking can be used to measure attention (*e.g.* does the film guide the viewer's attention as ex-

pected) or skin response can be used to measure arousal, often as a proxy for emotional impact or shockingness (Branje et al. 2014).

Achieving Control in Experiments

In any experiment, it is important to control for different factors that may cause the observed measurements. For evaluations of film, and particularly indirect evaluations, such control is particularly challenging. Increased control in experiments provides a number of benefits, for example: it reduces variance, increasing the statistical power; it allows us to better attribute measured effects to experimental manipulations (the things that are changed between conditions); and it helps insure repeatability of the experiments.

For comparison, we would like experiments to have the same conditions: everything should be the same, except for the aspect that we are trying to manipulate. For example, to compare editing techniques, the same conditions means that two techniques should be evaluated on the same contents, with the same choice of shots, lighting, 3-D animation or live actor performances, etc. This precludes comparison between different kinds of examples, for example, machinima film editing and feature film editing may have the effects of editing drowned out by the quality of the acting. It also means that comparison requires sharing of data (or systems to use on private data). Achieving similar conditions for films is also challenged by the nature of how aspects are intertwined: each live performance may be different in subtle ways, a lighting setup may be good for some camera angles and not others, etc. Keeping conditions similar also creates challenges with experimental design. A viewer watching the same (or similar) videos will have a different reaction to them. It is not possible to compare the experience of being told a story for the first time, and being told the same story a second time because the two experiences are ontologically different. Similarly, there may be effects of fatigue and boredom.

The work of Andrist *et al.* (Andrist et al. 2012) gives an example of the kinds of care required to address these issues. Experimental participants watched 4 videos, one for each different condition, and answered a questionnaire about each to measure their recall of information in the videos, as well as their impressions of the character presenting the information. For this study, four stories were developed, each designed to have equal amounts of information, and to be equally unfamiliar to the experimental participants. Four different characters were developed, similar enough in appearance to avoid effecting viewer impressions, but different enough to not appear repetitive. Randomization of presentation order to the experimental subjects was also used to control for fatigue effects.

Controlled experiments can be useful even in more direct assessment of editing. A first example of user study can be found in the work of Friedman and Feldman (Friedman and Feldman 2006), a knowledge-based approach which uses editing rules and principles collected from textbooks and from interviews with a domain expert. This work was evaluated with several example scenes from several TV genres. For evaluation, the authors showed their results together with manually edited movies to both experienced filmmakers and naive viewers. They conducted an informal evaluation by presenting several movies to the viewers and asking the audience to fill in a questionnaire. The main idea was to see if viewers, and especially filmmakers, could tell the difference between machine editing and expert human editing.

A second example is the work of Callaway *et al.* (Callaway et al. 2005). According to them, the cinematic expression is an arsenal of principles and conventions, which professional filmmakers learn, while naive viewers are usually not consciously aware of them. They consequently made the assumption that an evaluation involving experts would be more useful than an evaluation by naive users. They evaluated their system GLAMOUR, producing short video documentaries, by involving three professional filmmakers: a professional documentary director, a TV director with former experience in multimedia production, and a multimedia designer. The three experts were interviewed separately in a 3-step process. Each expert was first asked to watch all three pairs of videos in a random order. For each video, he was requested to assign an absolute score from 1 (terrible) to 10 (perfect). For each pair, he was then requested to directly compare the two videos on a scale from 1 (preference for default version) to 10 (preference for full cinematic version). The expert was finally asked to give some feedback on videos through an unstructured interview.

This study has shown some disagreement among the experts. Experts appeared to often be biased by their own particular styles and preferences, which makes numeric scores of little use. However, another conclusion of this study is that interviews might instead be more powerful means to assess the results and help improve editing systems.

Sharing datasets and codes

One of the biggest problems in evaluating editing system is certainly the lack of publicly available datasets. Indeed, the different options for evaluation require that the same dataset be made available to all methods for comparison purposes.

The benefits of sharing data sets and creating competition can be best illustrated in the area of computer vision. Most papers in object recognition in the early 2000's were dedicated to 4 object classes (bicycle, motorcycles, faces and cars) and shared datasets made it possible to measure progress towards solving those four problems (Ponce et al. 2006; Everingham et al. 2010). This in turn encouraged the computer vision community to address a more ambitious challenge of recognizing 101 object classes (Fei-Fei, Fergus, and Perona 2007). Today, the latest challenge encompasses more than 5000 objects classes (Fei-Fei 2010). Similarly the field of action recognition has developed at a very quick rate from two basic activities in the early 2000's (running and walking) to a dozen action classes (Weinland, Ronfard, and Boyer 2006) to 101 action classes (Soomro, Zamir, and Shah 2012) in just over a decade. In both cases, sharing datasets allowed the research community to advance at a faster rate by providing tools to measure progress.

One useful action for promoting future research in automatic film editing is therefore to create and distribute open datasets that the community can agree to use to evaluate

its own progress. This raises difficult issues such as format (should it be video with content annotation? animated 3d scenes?). This also raises the question of the number of scenes that would be useful, and on how much annotation should be provided on these scenes so as to be useful for any future editing system. Another question that comes up is how much complexity and variations should be offered.

In order to properly compare editing systems, all other things (story, acting, lighting, etc.) must be equal; there is a need to control everything that should not be manipulated by editors. As people will not be able to objectively compare the quality of two edits of the same story, a key element in properly designing an experiment is to provide different stories (if viewers have to watch three one-minute videos, then one should provide three different stories). Furthermore, in order not to introduce bias through differences in the stories engagement, then stories should be self-contained stories, and should also be equally challenging stories.

Conclusion

Indirect assessment of film editing techniques, by empirical studies measuring the effects of the films they produce, has a number of advantages. While it is challenging to create such experiments, it is possible to do. Shared data sets and codes would be helpful for this effort. Promoting this effort appears to be a task for which the WICED series of workshop is ideally suited.

Acknowledgments

Part of this work was supported by ERC Advanced Grant "Expressive".

References

Andrist, S.; Pejsa, T.; Mutlu, B.; and Gleicher, M. 2012. Designing effective gaze mechanisms for virtual agents. In *SIGCHI conference on Human Factors in Computing Systems*, 705–714. ACM.

Apter, J. 2014. The invisible art? *The New York Times, March 3 edition*.

Bloch, G. 1986. *Eléments d'une machine de montage pour l'audio-visuel*. Ph.D. Dissertation, Télécom Paris.

Branje, C.; Nespoil, G.; Russo, F.; and Fels, D. I. 2014. The effect of vibrotactile stimulation on the emotional response to horror films. *Computers in Entertainment* 11(1):5:1–5:13.

Callaway, C.; Not, E.; Novello, A.; Rocchi, C.; Stock, O.; and Zancanaro, M. 2005. Automatic cinematography and multilingual NLG for generating video documentaries. *Artificial Intelligence* 165(1):57–89.

Christian, D. B., and Young, R. M. 2004. Comparing cognitive and computational models of narrative structure. In *Proceedings of the 19th National Conference on Artifical Intelligence*, AAAI'04, 385–390. AAAI Press.

D'Ydewalle, G., and Vanderbeeken, M. 1990. Perceptual and Cognitive Processing of Editing Rules in Film. In *From eye to mind: Information acquisition in perception, search, and reading*.

Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision* 88(2):303–338.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2007. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding* 106(1):59–70.

Fei-Fei, L. 2010. Imagenet: crowdsourcing, benchmarking & other cool things. In *CMU VASC Seminar*.

Friedman, D., and Feldman, Y. A. 2006. Automated Cinematic Reasoning about Camera Behavior. *Expert Systems with Applications* 30(4):694–704.

Gleicher, M. 2012. Why ask why ? considering motivation in visualisation evaluation. In *BELIV*, 10:1–10:3. New York: ACM.

Graesser, A.; Lang, K.; and Roberts, R. 1991. Question answering in the context of stories. *Journal of Experimental Psychology: General* 120(3):254–277.

Harris, M. 2008. Which editing is a cut above? *The New York Times, January 6 edition*.

Jhala, A., and Young, R. M. 2009. Evaluation of intelligent camera control systems based on cognitive models of comprehension. In *Proceedings of the 4th International Conference on Foundations of Digital Games - FDG '09*, 327.

Kučera, D., and Haviger, J. 2012. Using Mood Induction Procedures in Psychological Research. *Procedia - Social and Behavioral Sciences* 69:31–40.

Magliano, J. P., and Zacks, J. M. 2011. The impact of continuity editing in narrative film on event segmentation. *Cognitive Science* 35(8):1489–1517.

Murch, W. 1986. *In the blink of an eye*. Silman-James Press.

O'Steen, B. 2009. *The Invisible Cut: How Editors Make Movie Magic*. Michael Wiese Productions.

Ponce, J.; Berg, T. L.; Everingham, M.; Forsyth, D. A.; Hebert, M.; Lazebnik, S.; Marszalek, M.; Schmid, C.; Russell, B. C.; Torralba, A.; Williams, C. K. I.; Zhang, J.; and Zisserman, A. 2006. Dataset issues in object recognition. In *Toward Category-Level Object Recognition (Sicily Workshop)*, 29–48. Springer Berlin Heidelberg.

Ponto, K.; Kohlmann, J.; and Gleicher, M. 2012. Effective replays and summarization of virtual experiences. *IEEE Transactions on Visualization and Computer Graphics* 18(4).

Smith, T. J. 2005. *An Attentional Theory of Continuity Editing*. Ph.D. Dissertation, University of Edinburgh.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. Ucf101: A dataset of 101 human action classes from videos in the wild. Technical report, CRCV-TR-12-01.

Weinland, D.; Ronfard, R.; and Boyer, E. 2006. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding* 104(2-3):249–257.