Copyright © 2023 ACM, Inc. The final publication is available at https://dl.acm.org/doi/10.1145/3587819.3590979 (see: https://authors.acm.org/author-resources/author-rights).

Semi-automatic mulsemedia authoring analysis from the user's perspective

Raphael Abreu
Federal Fluminense University
Niterói, Brasil
raphael.abreu@midiacom.uff.br

Douglas Mattos
Federal Fluminense University
Niterói, Brasil
douglas@midiacom.uff.br

Joel Santos
CEFET/RJ
Rio de Janeiro, Brasil
jsantos@eic.cefet-rj.br

George Guinea
Brunel University
London, United Kingdom
george.ghinea@brunel.ac.uk

Débora C. Muchaluat-Saade
Federal Fluminense University
Niterói, Brasil
debora@midiacom.uff.br

ABSTRACT

Mulsemedia (Multiple Sensorial Media) authoring is a complex task that requires the author to scan the media content to identify the moments to activate sensory effects. A novel proposal is to integrate content recognition algorithms into authoring tools to alleviate the authoring effort. Such algorithms could potentially replace the work of the human author when analyzing audiovisual content, by performing automatic extraction of sensory effects. Besides that, the semi-automatic method proposes to maintain the author subjectivity, allowing the author to define which sensory effects should be automatically extracted. This paper presents an evaluation of the proposed semi-automatic authoring considering the point of view of users. Experiments were done with the STEVE 2.0 mulsemedia authoring tool. Our work uses the GQM (Goal Question Metric) methodology, a questionnaire for collecting users' feedback, and analyzes the results. We conclude that users believe that the semi-automatic authoring is a positive addition to the authoring method.

CCS CONCEPTS

• **Human Computer Interaction** → **Interactive systems and tools**; • **Computer systems organization** → Other architectures; • **Computing Methodologies** → *Machine Learning*.

KEYWORDS

Semi-automatic authoring, sensory effects, user experiment, authoring tool

1 INTRODUCTION

In recent years, thanks to emerging technological advances in ubiquitous computing, there has been a resurgence of interest in increasing user immersion in virtual worlds by engaging more human senses. Devices such as scent emitters[1] or others that generate tactile sensations[2] have witnessed a proliferation. Additionally, there is a growing development of applications that stimulate other senses in conjunction with audiovisual content. These advances can lead to the creation of new experiences and open up opportunities for new ways of engaging users with multimedia content.

To define multimedia applications that explore other human senses the term mulsemedia (*Multiple Sensorial Media*) [14] was proposed. Unlike traditional multimedia applications, which are exclusively audiovisual (*i.e.*, vision and hearing), mulsemedia applications are those that involve, in addition to audiovisual content, one or more additional human senses (*e.g.*, touch and smell). Mulsemedia applications can also use sensing devices to identify environment and user states (*e.g.*, temperature and user reaction) and actuators to render sensory effects (*e.g.*, wind, fog and heat).

To create a mulsemedia application, the author needs to carefully inspect the audiovisual content to identify and annotate it with metadata defining a sensory effect at a given moment, its position, and specific attributes such as intensity. We call this process *authoring*. This manual authoring process is costly and misleading [1]. Therefore, one way to encourage the authoring of mulsemedia applications is to reduce the burden of manual authoring, especially by using intelligent systems that can automate the process of authoring sensory effects.

To accelerate the authoring of applications with sensory effects, several studies [2, 17, 21, 23, 25] proposed the integration of multimedia content analysis algorithms in the authoring of sensory effects. The basic idea is that algorithms replace the work of the human author when analyzing audiovisual content in search of information that may indicate the activation of a sensory effect. For example, camera movement indicating vibration [17], scene luminosity indicating light effects [23] or use of Deep Neural Networks (DNNs) for scene analysis [2] to automatically annotate sensory effects such as wind and heat.

[1] https://feelreal.com/
[2] https://teslasuit.io/

Although such techniques are powerful, there are limitations regarding their use to support mulsemedia authoring. Sensory effects have, in addition to their type and the moment of activation, specific characteristics that need to be automatically recognized, such as intensity and position. As discussed in [3], the authoring process is a highly creative task and fully automatic solutions may prevent the creative process or fail to meet the author's expectations. Additionally, since there is a myriad of possible inputs and outputs for methods of recognition, some way to provide interoperability is needed. Even so, considering the subjectivity of the authoring of effects, it is expected that authors can adapt the response of the recognition process to their preferences, *e.g.*, choose not to identify aromas.

To solve such challenges, in a previous work [3] we outlined a blueprint to develop a component that integrates content recognition on to existing mulsemedia authoring tools. The component acts as a plug-in to the existing software, enabling the use of content recognition software to perform the automatic annotation of sensory effects. This new component allows for configuration in accordance with the author preferences before and after the automatic annotation. Therefore, the author can fine-tune which sensory effect types should be recognized and which labels from the content recognition software might be associated with a sensory effect. This method is called **semi-automatic sensory effect authoring**. In [3] that component was integrated into STEVE 2.0 (*Spatio-TEmporal View Editor*) [13], a graphical authoring tool for mulsemedia applications. The component used one neural network as a recognition module and the results of the sensory effects extraction was compared with the annotation provided by the video dataset used. As an indication of the component efficacy, it was found a 61.4% match of the component annotation in relation to the one provided in the dataset.

While in [3] we demonstrated the technical capabilities of the component, it is yet to be understood if users view the addition of the component in an authoring tool as favorable. More importantly if they have the intent to use it in their authoring process. In this study, we aim to evaluate the effectiveness of the automatic extraction of sensory effects component in the multimedia authoring process from the perspective of users. To achieve this goal, we employed a combination of constructs (concepts or variables that are being studied to understand how technology is perceived, adopted and used by users) including the perceived usefulness and ease of use as outlined in the Technology Acceptance Model (TAM) [10]. Additionally, we evaluated the user's perception of the automatically extracted content, and the time required to use the component. It is important to note that our evaluation is specifically focused on this component within an authoring tool, rather than evaluating the tool as a whole. Our objective is to determine if the component can reduce the effort required for mulsemedia authoring. Thus, the main focus of the paper is to conduct a user-centered evaluation to test this hypothesis.

The remainder of the text is organized as follows. Section 2 presents background about content recognition and how it can be used to perform the authoring of sensory effects. Section 2.1 explains the tool used in this evaluation and how the semi-automatic authoring is performed. Section 3 presents related work to content recognition, authoring tools and automatic authoring of sensory effects. Section 4 presents our evaluation methodology and results. Finally Section 5 concludes our work presenting lessons learned and future work.

2 CONTENT RECOGNITION AND SEMI-AUTOMATIC AUTHORING

Content recognition is achieved by sending the media content to recognition software, which is software that employs algorithms capable of detecting objects (or concepts) in audiovisual media content. Those return a set of labels that indicate the description of objects (or concepts) at a given time in the media content. Deep Neural Networks (DNNs) have proved to be an effective method for analyzing image and video content according to [16, 24].

Figure 1 illustrates labels returned from a video recognition task with a DNN. In the figure, a 4-second video is shown and, for each second, a set of labels is presented. For brevity, only the most relevant 3 labels (top-3) are presented and their occurrence probabilities have been omitted. In the figure, we can see that the returned labels change as video content changes. At 1s from the start, the sun appears in the video and therefore the label `sun` starts to be returned.

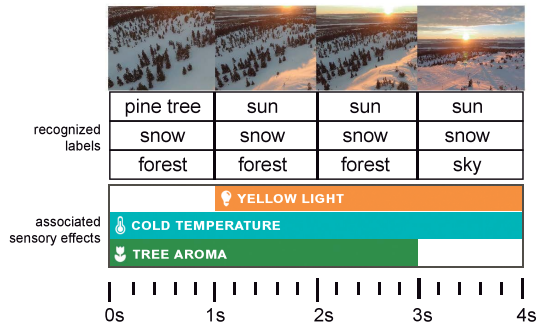

Figure 1: Sensory effect synchronization based on labels returned by DNN.

Labels returned from the recognition process can be associated with sensory effects such that, for example, whenever label `sun` occurs, there will be a light effect. Figure 1 also presents a timeline of sensory effect synchronization based on recognized labels. In the example, it is desired to synchronize labels `sun`, `snow` and `forest` with yellow light, cold temperature and tree aroma sensory effects, respectively.

As stated in [3], the main issue preventing machine-learning-based content analysis methods from being used for mulsemedia authoring is the lack of description standards dedicated to relating label naming with sensory effects. For example, to activate a wind effect we should use only the label `wind`, or a more complex description like `explosion` or `beach`. Another problem is that a DNN that was trained to classify content in daylight videos, once embedded into an authoring tool, may become unable to classify future content in darker videos. Thus, the recognition method has to be decoupled from the authoring tool. Furthermore, deciding where to place sensory effects is an often subjective decision-making process that involves an author's preference.

For such reasons as discussed previously, a more effective solution is to enable the author to select which DNN should be used to recognize sensory effects as well as which labels to be related to a given effect type. This tool was proposed as the *content-driven component* (CDC) in [3]. The CDC is a set of guidelines and an implementation to be incorporated into an authoring tool. With them, the tool can incorporate content recognition algorithms and provide a mechanism for adapting the response to annotate sensory effects on the timeline. The usage of the CDC with the STEVE 2.0 authoring tool, named STEVEML, will be discussed in the following section.

2.1 Semi-automatic Authoring in STEVE 2.0

The graphical interface of STEVE 2.0 can be seen in Figure 2. In the interface, the media repository at the upper left corner allows the author to import media objects into the graphic environment. In the upper center, we see the panel to edit the properties of the media objects and sensory effects. In the upper right, there is the preview screen for mulsemedia applications displaying their audiovisual content. The temporal view is presented at the bottom of the screen. This temporal view corresponds to an event-based timeline where nodes are synchronized using event-based causal relationships. These relationships and the entities that represent the mulsemedia application in STEVE 2.0 are defined by the MultiSEM [13] mulsemedia model.

From the media repository, the author can select a particular media object, drag it into the temporal view and create temporal relationships with other objects present in the timeline. To support sensory effects, STEVE 2.0 presents a list of sensory effect types above the temporal view so that authors can also drag a certain type of effect into the temporal view to create a new instance for the selected sensory effect. STEVE 2.0 allows the addition of wind, water spray, vibration, temperature, aroma, light, fog, flashlight, and the composite storm effect (rainstorm). The storm effect encompasses the effects of water spray, flashlight, and smoke.

The process of manually authoring sensory effects is carried out by dragging the sensory effect icons and placing them at the timeline. As soon as the author drags the icon to the timeline, a standard-duration sensory effect is inserted. The author can click on the effect icon in the timeline and change its properties, *e.g.*, its duration.

The process of semi-automatic authoring using STEVEML (the CDC implementation in STEVE 2.0) is carried out as follows. First, the author selects a media object and selects the option "AutoExtract Sensory Effects" in the STEVE mouse context menu. Then a pop-up window allows the author to select which sensory effect types should be recognized in the current media object. The pop-up window also allows the author to select which time slice of the media should be sent for content recognition. After the recognition, the corresponding sensory effect instances are added to the timeline.[3]

[3]We invite the reader to watch the accompanying video showcasing STEVEML in https://youtu.be/0OziKkuMeVQ

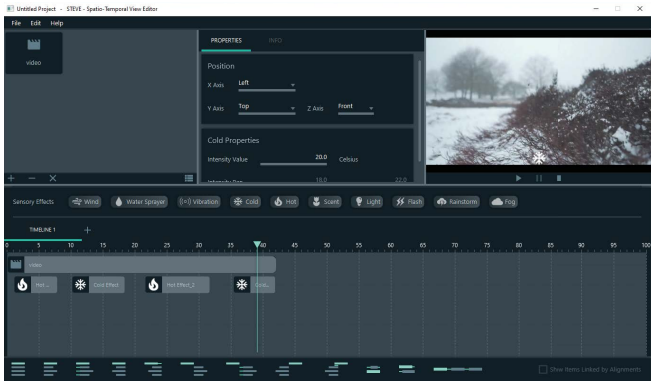

Figure 2: STEVE 2.0 graphical interface

3 RELATED WORK

As discussed in Covaci et al. [8], the quest for facilitating mulsemedia authoring has resulted in several authoring tools been developed by academia. One of the first is *SEVino* (*Sensory Effect Video Annotation*)[22]. In common with the surveyed tools, *SEVino* provides a graphical interface to the author that presents a video timeline to use as a basis for synchronizing sensory effects. The tool allows one to create time intervals that represent the duration of sensory effects. After the authoring phase, it generates MPEG-V-compliant descriptions indicating the temporal synchronization of sensory effects.

As pointed out by Waltl et al. [22], given the difficulty in authoring mulsemedia applications, an automatic form of authoring would encourage community adoption of such applications. A primary effort in this direction is the `autoExtraction` attribute in MPEG-V, which indicates whether extraction of a sensory effect is preferable. Although supported in the MPEG-V standard, it depends on the implementation of software capable of performing this automatic extraction. Tools supporting `autoExtraction` should perform it at run-time [22], *i.e.*, for the content already being played for the end-user. Thus, its temporal synchronization is completely automatic and independent of the application's author.

It is important to note that a fully automatic generation of sensory effects may be undesirable. After all, such authoring is an artistic process that depends on the preference of a human author to provide an enhanced user experience. Besides, fully automated proposals for authoring sensory effects have suffered negative repercussions from users in favor of human-generated ones. For instance, Lee et al. [18] report that authors of haptic effects disliked the completely automatic solution employed in the study. They see haptic authoring as a highly creative task and therefore believe it should be under author control. Thus, a better option for serving users and authors alike is to support sensory effect extraction at authoring time and give as much fine-tuning control to the author as possible. This is the approach adopted in our work.

The survey of Mattos et al. [20] reviews several mulsemedia authoring tools and proposals for representing sensory effects and their characteristics. The article intends to be a guide to develop better mulsemedia authoring tools and also outlines a set of desirable features for mulsemedia authoring tools - among them, that

an authoring tool should offer a graphical user interface approach that can guide authors in their production process. In particular, the authors outline the desirable feature for automatic extraction of sensory effects. That is, tools should allow authors to automatically extract sensory effects from audiovisual contents to enhance sensory effect annotation.

Kim et al. [15] and Danieau et al. [9] propose algorithms to extract sensory effects at runtime and at authoring time. Both approaches consist of using objective measurements based on image or sound processing to characterize information that enables sensory effects, such as pixel colors or loudness levels. The effects are added to the timeline of the authoring tool, which enables authors to fine-tune the results. One shortcoming of their approach is that the proposed algorithms are unable to identify complex elements in audiovisual content related to sensory effects (*e.g.*, beach, wind, rain, forest).

Amorim et al. [12] follow a different approach by employing *crowdsourcing* to gather the moments of activation of sensory effects. They also allow authors to fine-tune the time intervals of sensory effects indicated through *crowdsourcing*. The downside of [12] is the inherent cost and additional time needed to use a *crowdsourcing* platform. Our proposal resembles this work in the sense that it will also provide an indication of automatically-extracted sensory effects and enable the author to fine-tune the results. Apart from this, our work is aimed at integrating content analysis into existing authoring tools to automatically identify the moments of activation of sensory effects. This results in a faster solution without the additional cost of a *crowdsourcing* platform.

4 SEMI-AUTOMATIC SENSORY EFFECT AUTHORING EVALUATION

This paper presents an evaluation to validate our hypothesis that an automatic content recognition method can reduce the authoring effort using a mulsemedia authoring tool. We employed the Goal Question Metric (GQM) [4] approach to structure our evaluation. We may summarize GQM as follows: each defined goal has a set of questions that are answered using pre-established metrics. Each metric results in one or more numerical values. Moreover, GQM also defines the purpose and the perspective of each goal. The purpose defines the object of study and why we are analyzing it. The perspective defines a particular angle or aspect for evaluation and from whom that evaluation is given.

Regarding the aforementioned purpose of this evaluation, the goals are defined and presented in Table 1. The questions that adhere to these goals are defined in Table 2. In the rest of this section, we will further explain our Goals, Questions and Metrics used to validate our hypothesis.

Table 3 presents the metrics used to answer G1, G2, and G3 questions. Metrics PU (Perceived Usefulness) and PEOU (Perceived Ease of Use) follow the definition presented in the Technology Acceptance Model (TAM) [10, 11, 19]. According to TAM, the intent of the user to use a system is considered to be influenced by two major constructs, perceived usefulness and perceived ease of use. Perceived usefulness is defined as the degree to which the person believes that using the particular system would enhance her/his job performance. Whereas the perceived ease of use is defined as the degree to which the person believes that using the particular system would be free of effort [11]. In this paper, we have chosen to use a single question to measure PU and PEOU for several reasons. First, the focus of our study was the evaluation of a single functionality within an authoring tool, specifically the semi-automatic extraction of sensory effects, rather than evaluating the entire STEVE authoring tool. Using a single question allowed us to specifically assess how this functionality was perceived. Additionally, a single question is quicker and easier to administer than a multi-item scale, which was important for our study given the limited time and resources available. Furthermore, the simplicity of a single question was also important for our study as we were conducting research with a non-technical population. This ensured that all participants were able to understand and respond to the question easily.

In addition to the constructs outlined in TAM, we also proposed two new constructs for our evaluation. The first, called PAE (Perceived Authoring Effort), measures users' perception of the quality of the content produced by the automatic extraction process. Specifically, it assesses whether users found the annotations synchronized with the video presented. We chosen to evaluate this metric with two questions, Q3 and Q4 related to the same construct PAE. The idea was to have a more granular response about the reason for the user to accept the automatic extraction. The questions Q3 evaluates the perception of the user for the quality of the automated annotation, while Q4 evaluates the perception of the effort that the user would make based on the automatic annotation performed. Lastly, the second metric, ETD (Expected Task Duration), measures users' perception of the duration required to use the recognition module in the STEVE 2.0 multimedia authoring tool. These metrics

Table 1: Experiment Goals

Goals	Definition
G1	Analyse the perceived usefulness of semi-automatic mulsemedia authoring from the user's perspective.
G2	Analyse the perceived ease of use of semi-automatic mulsemedia authoring from the user's perspective.
G3	Analyse the perceived quality of the synchronization of automatic extraction from the user's perspective.

Table 2: Questions for Goals G1, G2 and G3

Goal	Question	Description
G1	Q1	Does the automatic extraction facilitate the authoring of sensory effects?
G2	Q2	Does the user perceive the automatic extraction functionality hard to use?
G3	Q3	Does the automatic extraction place sensory effects at different times than expected by the human author?
	Q4	Does the user perceive the need of a high authoring effort to re-synchronize sensory effects after automatic extraction?
	Q5	What is the response time of the automatic extraction functionality?

Table 3: Metrics for G1, G2 and G3 Questions

Metric	Description	*Question*
PU	*Perceived Usefulness* [10] refers to "the degree to which a person believes that using a particular system would enhance his or her job performance"	Q1
PEOU	*Perceived Ease of Use* [10] refers to "the degree to which a person believes that using a particular system would be free from effort"	Q2
PAE	*Perceived Authoring Effort* refers to it as the degree to which a person believes that the response from the automatic extraction would need effort to adapt to to their preferences	Q3 Q4
ETD	*Expected Task Duration*, measured in the reported duration of the authoring task with or without the automatic extraction	Q5

will be evaluated by aggregating the responses from users to the questionnaires.

4.1 Experimental Protocol

4.1.1 Users and Experiment Setup. Forty three (43) users participated in the experiments. Thirty five (35) were computer science students and nine (9) were from other areas, such as cinema, medicine, mathematics, physics, history, and law. In a pre-test questionnaire, the participants also reported how often they use a video editor application, 65,9% used occasionally, 27,3% never used and 6,8% frequently used a video editor. Only 34 participants completed all the necessary tasks.

The participants in the study conducted the experiments independently and remotely using a website with instructions. Prior to the experiments, an online presentation was provided to introduce the concept of multisensory applications and encourage participation. The main features of the authoring tool, STEVE, were then demonstrated through short videos to familiarize the participants with its functionality. The participants were provided with instructions on how to download and install STEVE and a test video, followed by tasks to be completed within the authoring tool. After completing each task, a post-test questionnaire was administered. All tasks were completed using the same test video.

4.1.2 Questionnaire. The post-test questionnaire consisted of four questions, labeled as Q1 to Q4 in Table 2, which were rephrased as positive or negative statements in order to eliminate bias in the original wording of the question and to facilitate the understanding of the questions. Participants had to answer them using a five-point Likert scale [6], ranging from 1 - Strongly disagree - to 5 - Strongly agree. Besides, one last question asked how much time the participant spent on the task. To answer it, the participant should indicate a numerical value representing the minutes taken to perform the task.

In this study, questions Q1 and Q2 pertain to the constructs of perceived usefulness (PU) and perceived ease of use (PEOU), respectively. We evaluated the internal consistency of questions Q3 and Q4, which pertain to the construct of perceived authoring effort (PAE), by using the raw mean inter-item correlation. This measure averages the correlation between all answers to the questions in the questionnaire. The resulting average inter-item correlation was found to be 0.39, which falls within the recommended range for this measure according to previous studies [5, 7].

4.1.3 Procedure. The experiment was divided into two tasks. The data used to evaluate our experiment goals were taken from Task 1 and Task 2 results.

In **Task 1** participants had to create a simple mulsemedia application with two sensory effects, hot and cold, and a video media object. The task goal was for participants to define the synchronization of both sensory effects with the video scenes without using the sensory effect extraction feature. Thus, users had to define the synchronization manually by dragging the effect type and media items into the STEVE temporal view, as presented in Figure 2.

In **Task 2** participants performed the same task as in Task 1, but now using the sensory effect extraction feature. After the tool presented the sensory effect extraction result, participants were asked to check if they agreed with the suggested temporal synchronization. To perform this task, participants had to select the automatic extraction of sensory effects feature in STEVE interface. Then, select only the hot and cold sensory effect types to be extracted. Finally, the user had to wait for the tool to update the timeline with the automatic annotated sensory effects, as it can be seen in Figure 3.

4.2 Results and Discussion

Figure 4 presents a box-plot of the authors' answers to the questions Q1, Q2, Q3 and Q4. In the Figure, the media is represented by a dashed line and the media is represented by a solid line. Table 4 shows the mean score for each question. As can be seen, the majority of the participants strongly agreed that the proposed functionality facilitates sensory effect authoring. The participants strongly disagreed that the functionality is difficult to use. Regarding the answers for question *Q3*, participants were not sure whether it would be necessary to change the auto-extract response. Finally, participants also disagreed that they would need a high authoring effort to fine-tune the automatically annotated effects.

Metric PU is computed as the mean score of question Q1. Metric PEOU is computed as the the mean score of question Q2. Metric PAE is computed as the the average of the mean score for questions Q3 and Q4. As responses were on the Likert scale ranging from 1 to 5, a mean value greater than 3 for one of those questions defines that users agree, on average, with the statement.

To evaluate our questionnaire responses, we employed a one sample T-test to determine whether the difference between the mean values from the questionnaire answers and a hypothesized value is statistically significant. In this testing our hypothesized value is the neutral/mid-point value of 3 (on a 5 point Likert scale) and a α of 0.05. To analyse the results we will present the sample mean (M) and the Standard Deviation (SD) of the population, along with the t statistic applied to the population and the p value.

Analyzing our questions, we can evaluate the following: On Q1 the users strongly agree (M = 4.76, SD = 0.6) that the automatic

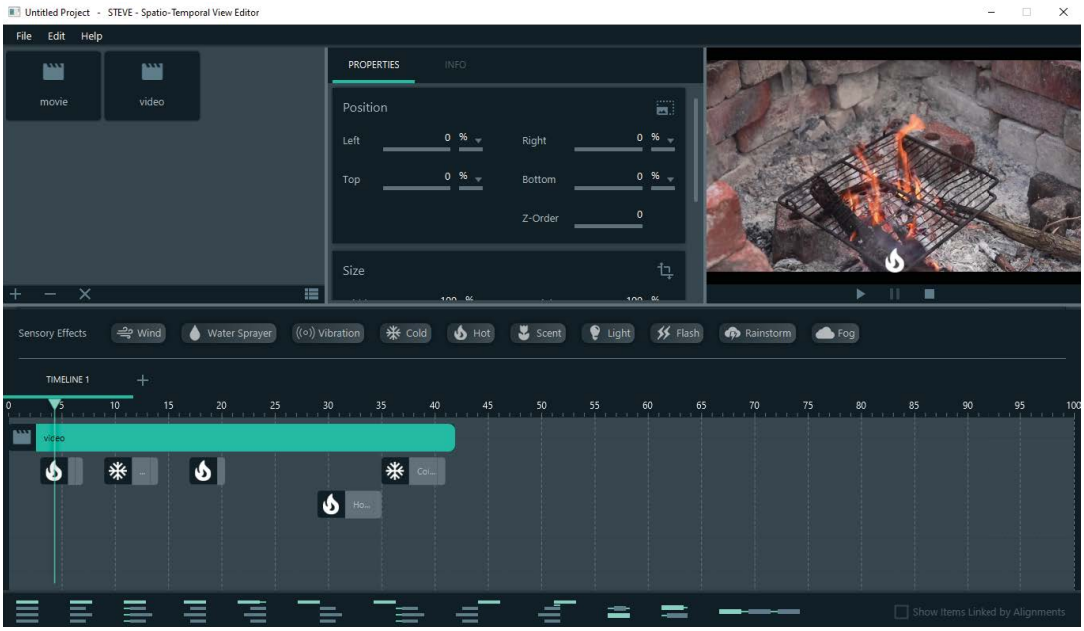

Figure 3: Automatic authoring result in STEVE for video.mp4

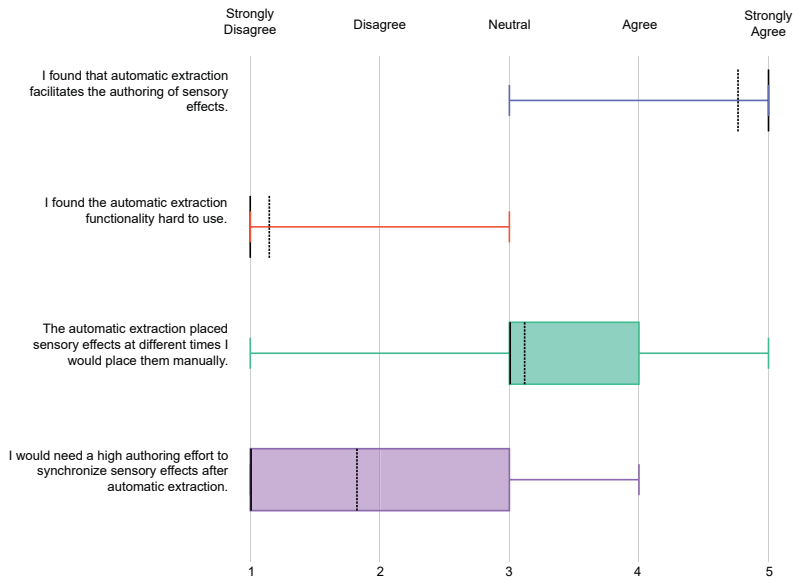

Figure 4: Answers from the questionnaire for Q1, Q2, Q3 and Q4

extraction facilitates the authoring of sensory effects $t(34) = 0.3$, $p < 0.001$. With this, we assign a value of 4.76 for PU and conclude that the users perceived the usefulness of the automatic extraction assistance in authoring sensory effects, achieving our goal G1.

On question Q2, the users strongly disagreed ($M = 1.15$, $SD = 0.43$) that the auto-extraction functionality was hard to use, $t(34) = 0.7$, $p < 0.001$. The results from Q1 and Q2 indicate that the users well perceived the automatic extraction functionality. With this, we assign a value of 4.85 for PEOU and conclude that the users perceived the ease of use of the automatic extraction functionality, this achieving our goal G2.

Values from Q3 indicate that participants were neutral ($M = 3.17$, $SD = 1.09$) about the question if automatic extraction differed from the expected annotation, $t(34) = 0.9$, $p = 0.186$. With $p > \alpha$ this difference is considered to be not statistically significant. One explanation for this result is that several participants perceived the

need to fine-tune the sensory effects after performing the automatic extraction (≈ 35%) while others disagreed with the sentence (≈ 20%). Finally, in Q4 the users agreed (M = 1.83 SD = 1.02) that task of adjusting the automatic extracted effects is low-effort, $t(34) = 0.13$, $p < 0.001$.

To create our metric PAE, we inverted the values of Q3 and Q4, as they were negative worded questions, resulting in the values 2.83 and 4.17, respectively. By averaging the inverted values of Q3 and Q4, we arrived at a metric for the construct of Perceived Automation Efficacy (PAE) of 3.5 (both questions are related to the same construct, as seen in Section 4.1.2). A higher value of this metric indicates a more favorable perception of the user towards the quality of content produced by the automatic extraction. While this result does not strongly validate our goal G3, Q4 provides an indication users had a favourable perception about the quality of the content produced by the automatic extraction. Nevertheless, further experiments are needed to better understand this relationship.

Table 4: Mean values for the answers of the questionnaire

Question	Mean value	SD.
Q1	4.76	0.6
Q2	1.15	0.43
Q3	3.17	1.09
Q4	1.83	1.02

Finally, we compute metric ETD from the self-reported time taken to complete the task. Figure 5 presents the results obtained. As it can be seen, participants spent on average 110.63 seconds on Task 2 with ≈ 93 SD. It is important to notice that no fine tuning is demanded in Task 2. For the sake of comparison, Figure 5 also shows the average time taken on Task 1. On that task, participants had to manually define the synchronization among the video and the sensory effects. They took, on average, 272.38 seconds to perform the task with ≈ 226 SD.

Comparing the results, we can observe that the average time taken to perform Task 2 is less than half of Task 1. This information by itself cannot validate goal G3 either, that is to confirm that semi-automatic authoring reduces the time spent during the authoring process. Besides, in Task 1, participants were interacting with the STEVE tool for the first time, what may increase the task duration. On the other hand, in Task 2, participants were not asked to perform any more editions. However, participants did respond in Q4 that they would not need a high authoring effort to synchronize sensory effects after the automatic extraction is performed.

4.2.1 Limitations. The evaluation presented here posed the following limitations. The value for the ETD metric was self-reported by the users, resulting in a unreliable measure of time taken to perform the task. Besides, task 2 was performed using the same video of task 1, therefore some users may have finished the task faster because the video was already known.

Besides the objective evaluation, the questionnaire also employed an open question to the user to report any findings or comments on the process. The responses mostly concerned the authoring tool functionalities and interface, not the automatic extraction feature. A few users mentioned problems with the authoring tool on their

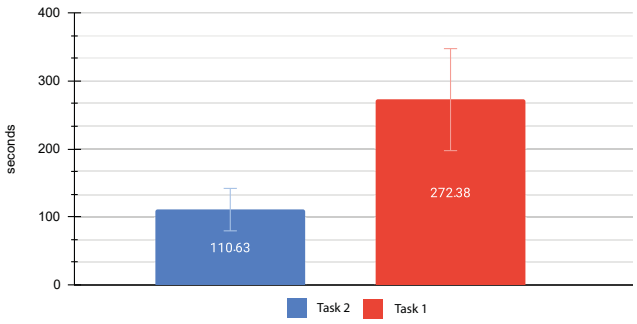

Figure 5: Self reported time taken to complete the task

systems, which led to their exclusion from the experiment. From the concluding users, one user mentioned "I saved the project as the automatic extraction synchronized by itself. It gives an start help, but I would have to sync myself if I wanted a perfect result. But, it is a good help tool to start the work.". This user's feedback corroborates our evaluation, since the automatic extraction was perceived as a good starting point for authoring. Another user mentioned "The automatic extraction wasn't completely off, but it failed to detect sensory effects in scenarios where the climate isn't strongly defined.". This can be viewed as a negative view of the system, however the method of semi-automatic authoring employed can change the recognition method to a better system that should be more tuned to the author's expectation.

5 CONCLUSION

This paper presented a user evaluation following a GQM structure. Therefore, three goals were defined to evaluate the perceived usefulness of sensory effect extraction used together with a mulsemedia authoring tool. To overview the experiment's results, most users successfully performed the automatic extraction. All users confirmed that the automatic extraction facilitates the authoring of sensory effects. The evaluation results indicate that the proposed recognition method is a viable alternative to reduce the authoring effort in a mulsemedia authoring tool.

Lessons learned from the preliminary evaluation raise the need for the creation of new metrics, with the aim to quantitatively evaluate the contribution of automatic extraction to the authoring workflow. An important metric would be to compute the amount of changes that the author would have to make after the automatic extraction is performed to adjust sensory effects in comparison with manual authoring. A second metric would be the time taken to perform the semi-automatic authoring in comparison with the manual authoring alone. Those metrics are left as future work.

6 ACKNOWLEDGMENTS

The authors wish to thank CAPES, CAPES Print, CNPQ, INCT-MACC and FAPERJ for the partial financing of this work.

REFERENCES

[1] Raphael Abreu and Joel dos Santos. 2017. Using Abstract Anchors to Aid The Development of Multimedia Applications With Sensory Effects. In *Proceedings of*

the 2017 ACM Symposium on Document Engineering (Valletta, Malta) *(DocEng '17)*. ACM, New York, NY, USA, 211–218. https://doi.org/10.1145/3103010.3103014

[2] Raphael Abreu, Joel dos Santos, and Eduardo Bezerra. 2018. A Bimodal Learning Approach to Assist Multi-sensory Effects Synchronization. In *International Joint Conference on Neural Networks* (Rio de Janeiro, Brazil) *(IJCNN '18)*. IEEE.

[3] Raphael Abreu, Douglas Mattos, Joel dos Santos, Gheorghita Ghinea, and Débora Muchaluat-Saade. 2000. Toward Content-Driven Intelligent Authoring of Mulsemedia Applications. 28, 1 (2000), 7–16.

[4] Victor R Basili. 1992. *Software modeling and measurement: the Goal/Question/Metric paradigm*. Technical Report.

[5] Stephen R Briggs and Jonathan M Cheek. 1986. The role of factor analysis in the development and evaluation of personality scales. *Journal of personality* 54, 1 (1986), 106–148.

[6] John Brooke. 1996. SUS: A Quick and Dirty usability scale. *Usability evaluation in industry, Chapter 21* (1996), 189–194.

[7] Lee Anna Clark and David Watson. 2016. Constructing validity: Basic issues in objective scale development. (2016).

[8] Alexandra Covaci, Longhao Zou, Irina Tal, Gabriel-Miro Muntean, and Gheorghita Ghinea. 2018. Is multimedia multisensorial?-a review of mulsemedia systems. *ACM Computing Surveys (CSUR)* 51, 5 (2018), 91.

[9] F. Danieau, J. Fleureau, P. Guillotel, N. Mollet, M. Christie, and A. Lécuyer. 2014. Toward Haptic Cinematography: Enhancing Movie Experiences with Camera-Based Haptic Effects. *IEEE MultiMedia* 21, 2 (Apr 2014), 11–21. https://doi.org/10.1109/MMUL.2013.64

[10] Fred D Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly* (1989), 319–340.

[11] Fred D Davis, Richard P Bagozzi, and Paul R Warshaw. 1989. User acceptance of computer technology: A comparison of two theoretical models. *Management science* 35, 8 (1989), 982–1003.

[12] Marcello Novaes de Amorim, Estêvão Bissoli Saleme, Fábio Ribeiro de Assis Neto, Celso A. S. Santos, and Gheorghiţă Ghinea. 2019. Crowdsourcing authoring of sensory effects on videos. *Multimedia Tools and Applications* (2019), 1–27.

[13] Douglas Paulo de Mattos, Débora C Muchaluat-Saade, and Gheorghita Guinea. 2020. An Approach for Authoring Mulsemedia Documents Based on Events. In *International Conference on Computing, Networking and Communications, 2020*. IEEE, 7.

[14] Gheorghita Ghinea, Christian Timmerer, Weisi Lin, and Stephen R. Gulliver. 2014. Mulsemedia : State of the Art, Perspectives, and Challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications* 11, 1s (2014), 1–23. https://doi.org/10.1145/2617994

[15] Sang Kyun Kim, Seung Jun Yang, Chung Hyun Ahn, and Yong Soo Joo. 2014. Sensorial information extraction and mapping to generate temperature sensory effects. *ETRI Journal* 36, 2 (2014), 224–231. https://doi.org/10.4218/etrij.14.2113.0065

[16] Yukhe Lavinia, Holly H. Vo, and Abhishek Verma. 2016. Fusion Based Deep CNN for Improved Large-Scale Image Action Recognition. In *2016 IEEE International Symposium on Multimedia (ISM)*. 609–614. https://doi.org/10.1109/ISM.2016.0131

[17] Jaebong Lee, Bohyung Han, and Seungmoon Choi. 2015. Motion effects synthesis for 4D films. *IEEE transactions on visualization and computer graphics* 22, 10 (2015), 2300–2314.

[18] Jaebong Lee, Bohyung Han, and Choi Seungmoon. 2016. Interactive motion effects design for a moving object in 4D films. In *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology*. ACM, 219–228.

[19] Nikola Marangunić and Andrina Granić. 2015. Technology acceptance model: a literature review from 1986 to 2013. *Universal access in the information society* 14 (2015), 81–95.

[20] Douglas Paulo De Mattos, Débora C Muchaluat-Saade, and Gheorghita Ghinea. 2021. Beyond multimedia authoring: On the need for mulsemedia authoring tools. *ACM Computing Surveys (CSUR)* 54, 7 (2021), 1–31.

[21] Thomhert S Siadari, Mikyong Han, and Hyunjin Yoon. 2017. 4D Effect Video Classification with Shot-Aware Frame Selection and Deep Neural Networks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 1148–1155.

[22] Markus Waltl, Benjamin Rainer, Christian Timmerer, and Hermann Hellwagner. 2013. An end-to-end tool chain for Sensory Experience based on MPEG-V. *Signal Processing: Image Communication* 28, 2 (2013), 136–150.

[23] Markus Waltl, Benjamin Rainer, Christian Timmerer, and Hermann Hellwagner. 2013. An end-to-end tool chain for Sensory Experience based on MPEG-V. *Signal Processing: Image Communication* 28, 2 (2013), 136–150.

[24] Matthew D. Zeiler and Rob Fergus. 2013. Visualizing and Understanding Convolutional Networks. *CoRR* abs/1311.2901 (2013). http://arxiv.org/abs/1311.2901

[25] Yuhao Zhou, Makarand Tapaswi, and Sanja Fidler. 2018. Now You Shake Me: Towards Automatic 4D Cinema. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7425–7434.