Assessing Usefulness, Ease of Use and Recognition Performance of Semi-Automatic Mulsemedia Authoring

RAPHAEL ABREU, MídiaCom Lab, Fluminense Federal University, Brazil
JOEL DOS SANTOS, CEFET/RJ, Brazil
GHEORGHITA GHINEA, Brunel University, United Kingdom
DÉBORA C. MUCHALUAT-SAADE, MídiaCom Lab, Fluminense Federal University, Brazil

Mulsemedia (Multiple Sensorial Media) authoring poses a considerable challenge as authors navigate the intricate task of identifying moments to activate sensory effects within multimedia content. A novel proposal is to integrate content recognition algorithms that use machine learning (ML) into authoring tools to alleviate the authoring effort. As author subjectivity is very important, it is imperative to allow users to define which sensory effects should be automatically extracted. This paper conducts a twofold evaluation of the proposed semi-automatic authoring. The first is from a user perspective within the STEVE 2.0 mulsemedia authoring tool, employing the Goal-Question-Metric (GQM) methodology and a user feedback questionnaire. Our user evaluation indicates that users perceive the semi-automatic authoring approach as a positive enhancement to the authoring process. The second evaluation targets sensory effect recognition using two different content recognition modules, quantifying their automatic recognition capabilities against manual authoring. Metrics such as precision, recall, and F1 scores provide insights into the strengths and nuances of each module. Differences in label assignments underscore the need for ML module result combination methodologies. These evaluations contribute to a comprehensive understanding of the effectiveness of sensory effect recognition modules in enhancing mulsemedia content authoring.

CCS Concepts: • **Human-centered computing** → *Usability testing*; • **Applied computing** → **Hypertext / hypermedia creation**; • **Information systems** → **Multimedia information systems**.

Additional Key Words and Phrases: 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*) [17] was proposed. Unlike traditional multimedia applications, which are exclusively audiovisual (*i.e.*,

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

Authors' addresses: Raphael Abreu, MídiaCom Lab, Fluminense Federal University, Niterói, Brazil, raphael.abreu@midiacom.uff.br; Joel dos Santos, CEFET/RJ, Rio de Janeiro, Brazil, jsantos@eic.cefet-rj.br; Gheorghita Ghinea, Brunel University, London, United Kingdom, george.ghinea@brunel.ac.uk; Débora C. Muchaluat-Saade, MídiaCom Lab, Fluminense Federal University, Niterói, Brazil, debora@midiacom.uff.br.

ACM 1551-6865/2024/8-ART
https://doi.org/10.1145/3689640

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, 21, 27, 29, 32] have 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 [21], scene luminosity indicating light effects [29] or use of Deep Neural Networks (DNNs) for scene analysis [2] to automatically annotate sensory effects such as wind and heat.

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 [4], 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 aroma effects. Furthermore, integrating results from different recognition modules is a significant challenge. Variations in label assignments and the differing strengths and weaknesses of each module make it difficult to present unified outcomes. Effective methodologies for combining outputs from diverse ML modules need to be explored to achieve more robust and accurate sensory effect recognition in multimedia content.

To solve such challenges, in previous work [4] 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 configuration in accordance with 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 [4] that component was integrated into STEVE 2.0 (*Spatio-TEmporal View Editor*) [15], a graphical authoring tool for mulsemedia applications. The component used one content recognition service as a recognition module and the results of the sensory effects extraction was compared with the annotation provided by the video dataset used. To indicate the component's efficacy, an average of 61.4% match was found between the component annotation and the annotations provided in a manual dataset.

While [4] demonstrated the technical capabilities of the component, its reception among users and integration into an authoring tool remain unexplored. This paper is an extended version of [3] addressing those gaps through a twofold evaluation. Firstly, from a user-centered perspective, the study employs constructs such as perceived usefulness and ease of use, drawing from the Technology Acceptance Model (TAM), to assess users' views on the proposed semi-automatic authoring approach within the STEVE 2.0 mulsemedia authoring tool. This first study was initially discussed in [3]. In the light of the responses of the user evaluation, this paper also presents a performance-oriented evaluation, which explores sensory effect recognition using two recognition modules,

aiming to quantify their automatic recognition capabilities against manual authoring. This second evaluation is an original contribution of this work.

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 3 presents related work to content recognition, authoring tools and automatic authoring of sensory effects. Section 4 explains the tool used in this evaluation and how the semi-automatic authoring is performed. Section 5 presents our user evaluation methodology and results. Section 6 presents our comparative evaluation of two recognition modules and results. Finally, Section 7 concludes our work presenting lessons learned and future work.

2 CONTENT RECOGNITION AND SEMI-AUTOMATIC AUTHORING

Content recognition is achieved by sending media content to recognition software, which is software that employs algorithms capable of detecting objects (or concepts) in audiovisual media content. These 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 [20, 31].

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. For example, at 1s from the start, the sun appears in the video and therefore the label sun starts to be returned.

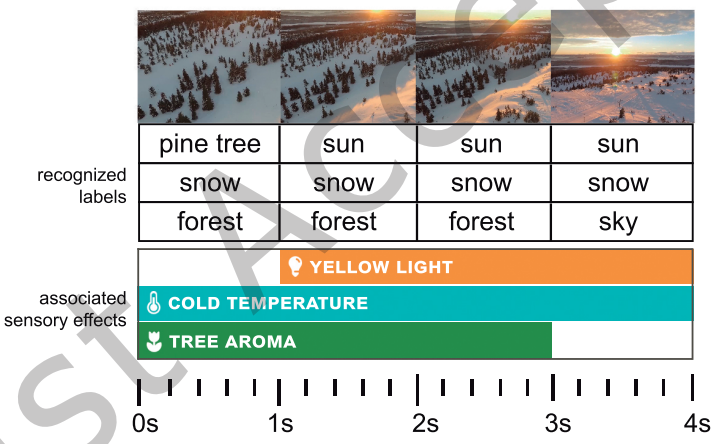

Fig. 1. Sensory effect synchronization based on labels returned by DNN.

Labels returned from the recognition process can be associated with sensory effects such that, for example, whenever label sun occurs, there will be a yellow 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 light, cold and aroma sensory effects, respectively.

As stated in [4], 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 often a 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 relate to a given effect type. This tool was proposed as the *content-driven component* (CDC) in [4]. 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.

3 RELATED WORK

As discussed in Covaci et al. [9], 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*) [29]. 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. [29], 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 [29], *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. [22] 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. [24] 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 - amongst 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. [19] and Danieau et al. [10] 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. [13] 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 [13] is the inherent cost and additional time needed to use a *crowdsourcing* platform. Our proposal resembles this work in the sense that it also provides 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.

Another tool for mulsemedia authoring is the STEVE 2.0 authoring tool [14, 15]. In STEVE 2.0, sensory effects can be synchronized with various traditional media (audio, image, and text) and not just with a single video. STEVE 2.0 also allows the author to create and synchronize sensory effects without the need for one main video or audio content to guide the application. Among the tools mentioned, only STEVE 2.0 was available for modifying the code to integrate our proposed semi-automatic authoring approach. The usage of the CDC with the STEVE 2.0 authoring tool, named STEVEML, will be discussed in the following section.

4 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 [15] 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 an 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]

5 SEMI-AUTOMATIC SENSORY EFFECT AUTHORING EVALUATION

We carried out a first evaluation, published in [3], 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) [5] 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.

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

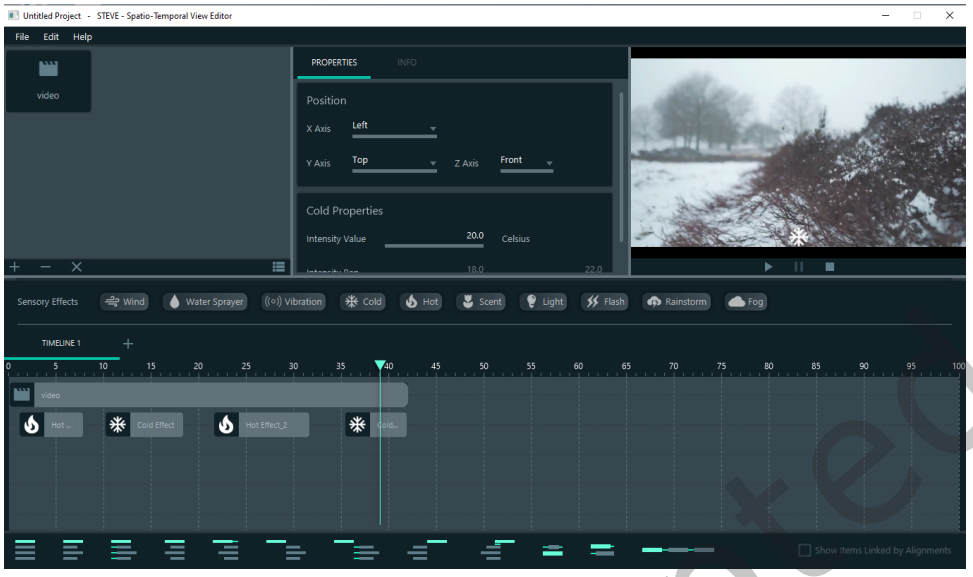

Fig. 2. STEVE 2.0 graphical interface

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 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 3 presents the metrics used to answer questions G1, G2, and G3 . Metrics PU (Perceived Usefulness) and PEOU (Perceived Ease of Use) follow the definition presented in the Technology Acceptance Model (TAM) [11, 12, 23]. 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 [12].

In this work, 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

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?

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 easily respond to the question.

While TAM traditionally employs multiple items to measure these constructs, in this study, we opted for a single, focused question for each. This decision to utilize single-item measures for PU and PEOU was constrained by the study's integration within a broader user experience assessment of the STEVE multimedia authoring tool, which included the use of the SUS and TAM questionnaires to evaluate the authoring tool. Our focus was on a single functionality within a larger authoring tool (the semi-automatic extraction of sensory effects), rather than the tool as a whole. A single, targeted question for each construct allowed us to efficiently assess perceptions of this specific feature. Additionally, the single-item format streamlined questionnaire completion, respecting participants' time.

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 presented video. 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. Question 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 will be evaluated by aggregating user responses to the questionnaires.

5.1 Experimental Protocol

5.1.1 Users and Experiment Setup. Forty four (44) 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. One participant reported being unable to finish due to the automatic extraction feature not responding. We will investigate and address this tool malfunction as part of our future work.

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

Table 3. Metrics for G1, G2 and G3 Questions

Metric	Description	*Question*
PU	*Perceived Usefulness* [11] 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* [11] 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

instructions on how to download and install STEVE and a test video, followed by tasks to be completed within the authoring tool. Upon completing the first task, users were asked to self-report the time taken. Following the second task, which utilized the same test video, a last questionnaire was administered with reported time taken to conclude task 2 and additional questions.

5.1.2 Questionnaire. The last 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 [7], 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 [6, 8].

5.1.3 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, participants had to wait for the tool to update the timeline with the automatic annotated sensory effects, as it can be seen in Figure 3.

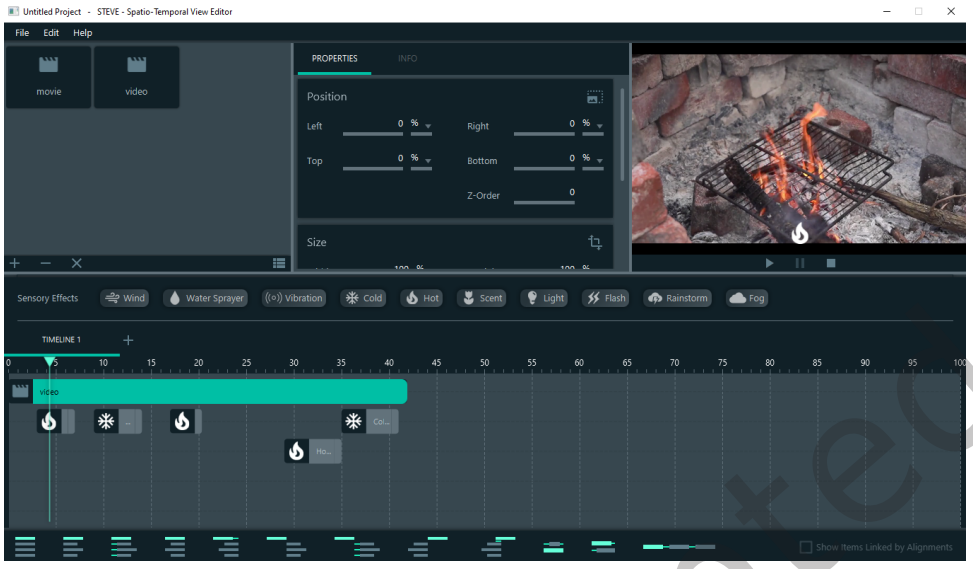

Fig. 3. Automatic authoring result in STEVE for video.mp4

5.2 Results

Figure 4 presents a box-plot of the authors' answers to questions Q1, Q2, Q3, and Q4, with the mean represented by a dashed line and the median by a solid line. Table 4 shows the mean score and standard deviation (SD) for each question. The majority of participants strongly agreed that the proposed functionality facilitates sensory effect authoring (Q1) and strongly disagreed that the functionality is difficult to use (Q2). Responses to Q3 were more neutral, indicating uncertainty about whether changes to the auto-extracted responses would be necessary. In Q4 the users tend to agree that the task of adjusting the automatic extracted effects is low-effort. However, responses for Q4 show a greater spread.

Table 4. Mean values and standard deviation (SD) 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

To ensure the internal consistency of our questionnaire items, we calculated Cronbach's alpha, a measure of reliability that indicates how well a set of items measure the same underlying construct. We obtained a Cronbach's alpha of 0.655. It is important to note that one question was reversed during this calculation due to its negative wording. While there is no single acceptable value for Cronbach's alpha, a value of 0.70 is often cited as a rule of thumb for good internal consistency [25]. However, lower values can be acceptable in exploratory research or when measuring complex constructs [28]. This suggests that our questionnaire items have moderate internal consistency in our study, indicating that they are generally measuring the same underlying construct.

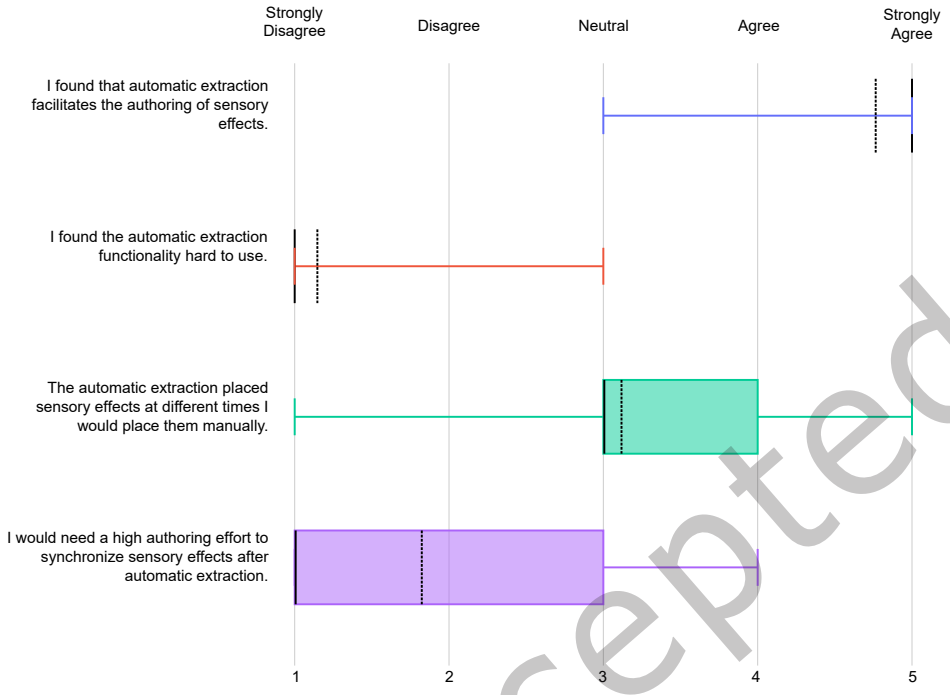

Fig. 4. Answers from the questionnaire for Q1, Q2, Q3 and Q4

To evaluate our questionnaire responses, we employed a one-sample Wilcoxon signed-rank test to determine if the median response for each question was significantly different from the neutral/mid-point value of 3 (on a 5-point Likert scale). Given that the Wilcoxon signed-rank test traditionally tests against a hypothesized median of 0, we normalized our data by subtracting 3 from each response, effectively shifting the neutral point to 0. We adopted a significance value (α) of 0.05. To analyze the results, we present the sample median (Mdn), which will be shown in its normalized form (where negative values represent responses below 3 on the original scale), and the Wilcoxon signed-rank test statistic (W), along with the associated p-value.

After analyzing our questions, we can evaluate the following: On Q1 the users strongly agree (M = 4.76, SD = 0.6) that the automatic extraction facilitates the authoring of sensory effects (Mdn = 2, W = 261.5, $p < 0.001$). On question Q2, the users strongly disagreed (M = 1.15, SD = 0.43) that the auto-extraction functionality was hard to use (Mdn = -2, W = 40, $p < 0.001$).

Values from Q3 indicate that participants were neutral. The distribution of responses for this question was not significantly different from 0 (Mdn = 0, W = 100, p = 0.0826). Finally, in Q4 the users agreed that the task of adjusting the automatic extracted effects is low-effort (Mdn = -2, W = 3.5, $p < 0.001$).

Finally, we compute metric ETD from the self-reported time taken to complete the task. Figure 5 presents the results obtained. As can be seen, participants spent on average 110.63 seconds on Task 2 with $\approx$ 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 $\approx$ 226 SD.

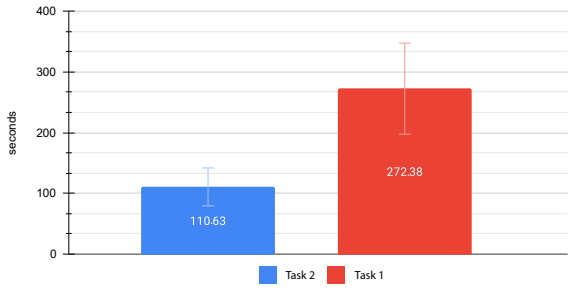

Fig. 5. Self reported time taken to complete the task

5.3 Discussion

In this section, we analyze the metrics derived from user responses and discuss how they address the evaluation goals and broader implications for the automatic extraction feature.

The PU metric, computed as the mean score of question Q1, yielded a value of 4.76. This high score indicates that users perceive automatic extraction assistance as highly useful, thereby achieving our goal G1. The PEOU metric, derived from the mean score of question Q2, resulted in a value of 4.85. This suggests that users found the automatic extraction functionality easy to use, successfully meeting our goal G2.

The PAE metric is calculated by averaging the mean scores of questions Q3 and Q4, which were reverse-scored due to their negative wording. This resulted in values of 2.83 and 4.17 for Q3 and Q4, respectively, yielding an overall PAE score of 3.5. A score above 3 suggests a favorable perception of the quality of content produced by the automatic extraction, although this does not strongly validate our goal G3. Notably, Q3 indicates that while some users needed to fine-tune the sensory effects after extraction (approximately 35%), others did not perceive this necessity (approximately 20%).

Finally, we compute metric ETD from the self-reported time taken to complete the task. We observed that the average time taken to complete Task 2 (with automatic extraction) was less than half the time required for Task 1 (manual authoring). While this suggests potential time-saving benefits of the automatic extraction feature, it does not fully confirm our goal G3. Factors such as participants' first-time interaction with the STEVE tool during Task 1 and the absence of further edits in Task 2 might have influenced the results. Additionally, responses to Q4 indicated a generally low perceived authoring effort after automatic extraction, supporting the notion that the tool facilitates the authoring process without requiring extensive edits.

Overall, the results highlight that users find the automatic extraction feature both useful and easy to use. However, the varying responses to Q3 and Q4 suggest that the system may not perfectly meet all users' expectations regarding automatic sensory effects placement and the required editing effort. This discrepancy can be attributed to the subjective nature of authoring. While some authors were satisfied with the automatically extracted sensory effects, others felt the need to extensively edit them to align with their individual intentions and creative vision. This highlights the need for a balance between automation and user control, ensuring that the tool assists the authoring process without stifling creativity.

5.4 Limitations

A key limitation of our user evaluation was the reuse of the same video in both tasks, which may have artificially inflated performance in the second task due to user familiarity. This design choice was constrained by the study's integration within the broader evaluation of the STEVE multimedia authoring tool. Additionally, the assessment relied on users' self-reported task completion times (ETD) as the tool did not allow for the capture of usage logs

to objectively track authoring behavior after the automatic extraction process, thus it may not fully reflect actual time spent by users. Consequently, these factors hinder a clear assessment of the automatic extraction's true impact on authoring efficiency and indicates the need for a more robust evaluation methodology in future studies.

It is important to note that relying on single-item measurements, such as the self-reported ETD, is not considered best practice in many research contexts [16]. This approach limits our ability to identify inconsistent or unreliable users, a particularly crucial aspect in remote studies.

One limitation of this study's statistical analysis was its reliance on a one-sample Wilcoxon signed-rank test to evaluate questionnaires. While appropriate for ordinal data, this approach may not fully capture the nuances of individual user experiences. A more comprehensive analysis, such as a mixed-effects model [26] accounting for user-level variability, could potentially offer deeper insights into the factors influencing user satisfaction. Additionally, incorporating a larger and more diverse dataset could enhance the generalizability and statistical power of the findings.

Moreover, our obtained Cronbach's alpha of 0.655, while acceptable for exploratory research, is below the commonly recommended threshold of 0.70, suggesting that there might be some room for improvement in the internal consistency of our questionnaire.

Furthermore, it is important to acknowledge that the participant pool primarily consisted of male computer science students. This demographic may not fully represent the diverse range of potential users for this tool, such as content creators. Their varying levels of technical expertise and specific use cases could significantly impact the perceived usefulness and effectiveness of the automatic extraction feature.

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 systems, which led to their exclusion from the experiment. Among users who did complete the experiment, one mentioned "I saved the project as the automatic extraction synchronized by itself. It gives a starting 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.

These comments also emphasize the importance of addressing recognition limitations and refining the accuracy of the automatic extraction system, further underscoring the motivation for incorporating additional recognition modules. This need arises from the recognition modules' varying capabilities in capturing the complexity of sensory effects, especially those that may be more intricate or context-dependent. Thus, this paper extends [3] with a new evaluation discussed in the following section.

6 COMPARATIVE EVALUATION OF SENSORY EFFECT RECOGNITION MODULES

In our user experience evaluation presented in Section 5, the recognition module used in STEVEML was the content recognition API called Clarifai[4], a cloud-based DNN service for video recognition. It used the *general* recognition model which can return over 11,000 different labels. As the STEVEML method can be adapted to use other recognition modules, we sought to compare the original Clarifai response with another module. In this pursuit, we introduce an extended evaluation that uses a new recognition module based on Amazon Web Services (AWS), specifically AWS Rekognition[5]. The AWS Rekognition module we employed uses a *general* recognition

[4]https://clarifai.com

[5]https://aws.amazon.com/rekognition/

model. Just like Clarifai, AWS Rekognition is a cloud-based DNN service that analyzes images and returns a set of labels.

The experiment aims to compare STEVEML automatic recognition considering both modules in relation to manual authoring of sensory effects using the sensory effects *dataset* introduced in [30]. The dataset consists of videos and related annotations according to the MPEG-V standard, focusing on sensory effects such as wind, light, and vibration. For this evaluation, we considered only the vibration effects and selected from the dataset the *Action* subset containing 38 videos, excluding three videos without manual annotations and two representing animations. The remaining 35 videos, ranging from 6 to 135 seconds, were used for this evaluation.

This evaluation procedure is a follow-up of the approach outlined in [4]. Our evaluation extends it with the comparison between modules. STEVEML enables the author to associate the recognition of a label with a sensory effect activation, thus signalling that a specific video segment can be synchronized with that particular sensory effect. Consequently, this evaluation aims to determine which manually annotated sensory effects were also identified by the modules. For the Clarifai module, the labels associated with the vibration effect include *calamity*, *motion*, and *action*. Conversely, for the AWS module, the labels associated with vibration effects encompass *Explosion, Weapon*, and *Fighting*.

6.1 Results and Discussion

Figure 6 illustrates the ratio of true positive matches between automatically authored sensory effects (both Clarifai and AWS) and manual authoring for each analyzed video[6]. The match is defined as an intersection between a given automatic annotation and one or more manual annotations. As highlighted in [1], it is posited that up to one second after the video scene, the vibration effect remains perceptually synchronous for users. Given this insight, we considered an intersection feasible within one second before or after the conclusion of manual vibration annotation.

As shown in Figure 6, both recognition methods exhibited considerable variation for some videos. The average true positive match rate for AWS was 52.80%, whilst for Clarifai this was 61.34%. Notably, Clarifai generally exhibits higher precision, while the AWS module achieves individual successes in certain videos. For instance, in videos 30 and 31, where Clarifai failed to identify any relevant concepts, AWS successfully recognized "weapon," which aligned with the manually authored content. Similarly, in video 8, AWS identified "smoke" (associated with "explosion") and aligned it with the manual annotation, despite Clarifai only identifying a generic "action". However, there were instances where Clarifai's broader recognition of "action" or "motion" (as in video 35) did not directly align with the manual authoring, while AWS successfully identified "weapon." While the percentage of true positives indicates that the module can keep up with manual authoring automatically, it cannot depict the complete picture of the modules' recognition. For a more comprehensive analysis, each module's ability to recognize false positives (sensory effects where humans marked none) is crucial. The ratio of false positives indicates if a module is performing poorly and could potentially hinder the author by introducing numerous errors that need correction. To provide a thorough evaluation of recognition performance, precision and F1 scores were computed for both the Clarifai and AWS modules.

Equation 1 defines the precision metric, which quantifies the accuracy of predictions. It is computed as the proportion of true positive instances relative to the aggregate of both true positive and false positive instances. As depicted in Figure 7, the precision scores for the Clarifai and AWS modules are compared. The precision scores for AWS and Clarifai modules averaged 0.41 and 0.54, respectively.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{1}$$

[6]Readers are invited to explore the names of the videos in Appendix A

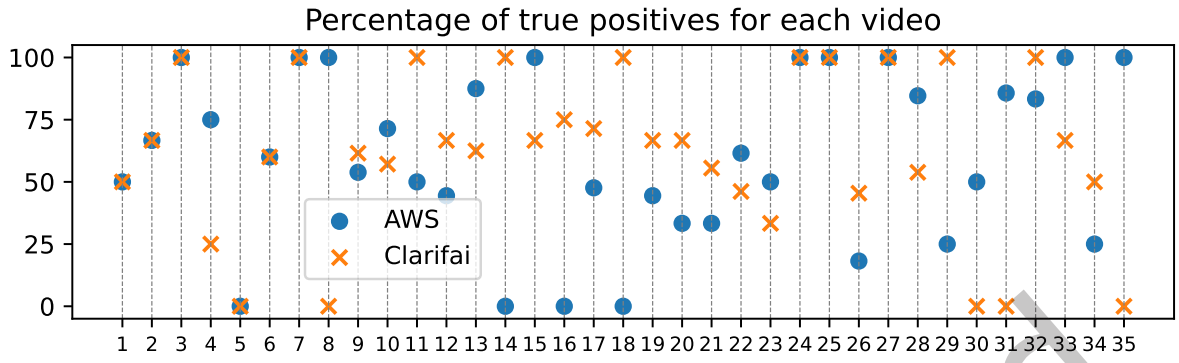

Fig. 6. Percentage of true positive matches of automatic annotation (Clarifai and AWS) with manual annotations for each video of the dataset.

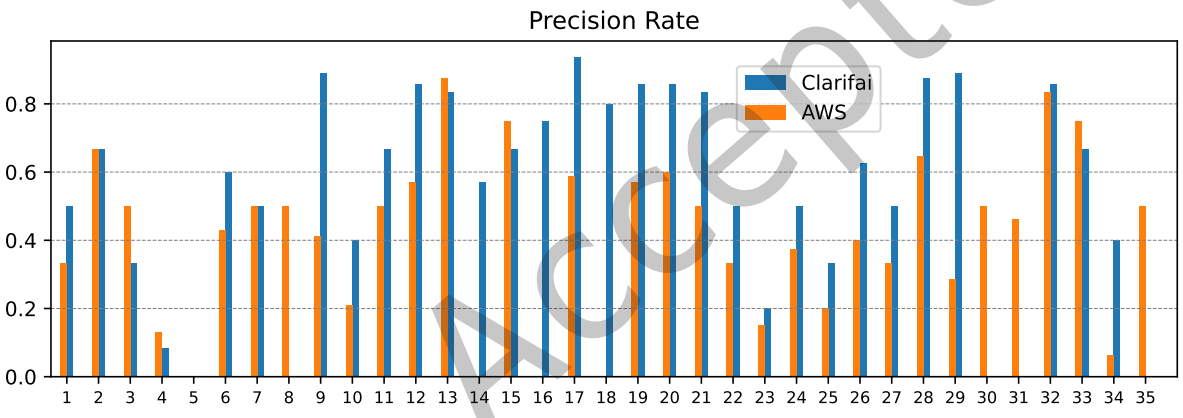

Fig. 7. Precision scores for Clarifai and AWS modules.

The F1 score is a robust measure that combines precision and recall, offering a balanced view of a model's performance. Precision, as defined in Equation 1, assesses the proportion of true positives among all positive predictions. Recall, detailed in Equation 2, measures the model's capability to identify all pertinent instances. The F1 score, articulated in Equation 3, is the harmonic mean of precision and recall, providing a single metric that balances both sensitivity (recall) and specificity (precision).

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{2}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

A higher F1 score indicates a more balanced and effective model, as it suggests an optimal balance between precision and recall. Figure 8 illustrates the F1 scores for the Clarifai and AWS modules, with Clarifai averaging an F1 score of 0.59 and AWS at 0.50.

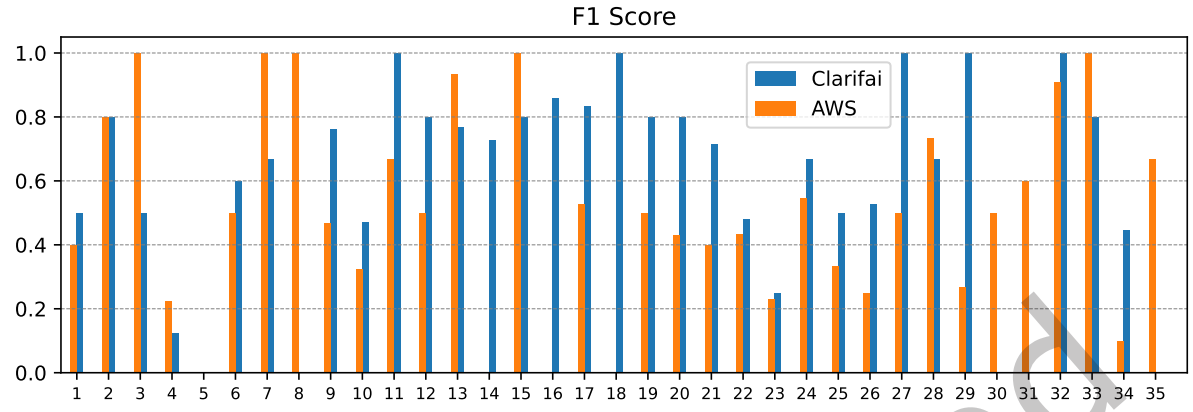

Fig. 8. F1 scores for Clarifai and AWS modules.

As Figure 8 highlights, both Clarifai and AWS modules exhibited instances of false positives, wherein the number of automatic annotations surpassed manually annotated effects. To exemplify this phenomenon, Figure 9 provides a comparative analysis of manual and automatic annotations in video number 28 (*babylonad*), incorporating both the Clarifai and AWS modules. Additionally, for the AWS module, the accuracy rate was approximately 84%, with 11 true positives (TP), 2 false negatives (FN), and 6 false positives (FP). In contrast, the Clarifai module achieved an accuracy rate of 53%, with 7 true positives, 6 false negatives, and 1 false positive.

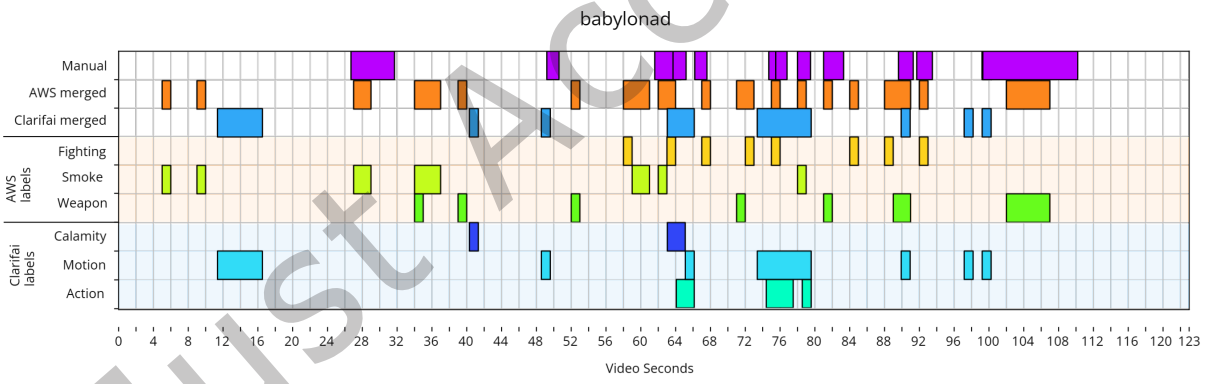

Fig. 9. Recognition on a timeline for the video *babylonad*.

While both Clarifai and AWS successfully recognized labels associated with sensory effects, such as categorizing explosions as '*calamity*', there were instances where effects present in manual annotations were overlooked. Figure 9 highlights differences in label assignments between Clarifai and AWS. Common labels related to the vibration effect, like *calamity, motion*, and *action*, are acknowledged by both modules, but variations exist in specific label assignments. For example, Clarifai may identify a particular effect where AWS does not, and vice versa. Notably, despite these differences, Clarifai and AWS exhibit complementary findings, as illustrated in Figure 9. This complementarity suggests that AWS and Clarifai can enhance the overall recognition process when used together, with each excelling in recognizing certain effects.

An open challenge that emerges is the efficient integration of results from different recognition modules. The variations in label assignments, coupled with differing strengths and weaknesses of each module, present a challenge in aggregating and presenting unified outcomes. Addressing this challenge involves exploring methodologies for combining outputs from diverse ML modules to achieve a more robust and accurate sensory effect recognition in multimedia content. Besides, it is important to adapt the system more to the author's expectations and intentions. This could be achieved by incorporating user interface feedback mechanisms to empower authors to refine the selection of labels for recognition (instead of relying solely on a pre-defined dictionary) and establish personalized rules for integrating multiple recognition models. For instance, users could specify how to handle scenarios where different models generate conflicting results, such as one module identifying "hot" effects while another identifies "cold" effects. Addressing these challenges in combining outputs from diverse recognition models represents an important avenue for future research, with the goal of developing more sophisticated and adaptable mulsemedia authoring systems.

7 CONCLUSION

This paper has presented a comprehensive evaluation of sensory effect extraction within a mulsemedia authoring tool, encompassing both user-centric and performance-oriented assessments. The user evaluation, structured following the Goal-Question-Metric (GQM) framework, aimed to gauge the perceived usefulness of automatic sensory effect extraction. The results revealed that the majority of users successfully executed the automatic extraction process, confirming its facilitating role in mulsemedia authoring. The findings suggest that the proposed recognition method serves as a viable alternative to alleviate the authoring burden in mulsemedia content creation.

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 lesson that will be implemented in future studies is the objective tracking of task duration inside the authoring tool, via logs. Another 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 further 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. Given that Cronbach's alpha is sensitive to the number of items, future research should also consider expanding the questionnaire to include more items assessing the same construct. This could potentially increase the reliability of the measure and provide a more comprehensive understanding of the internal consistency of the user experience assessment.

In addition to the user-centric evaluation, we conducted a performance-oriented assessment focusing on the quantitative aspects of sensory effect recognition. Comparing two recognition modules, namely Clarifai and AWS, we examined their ability to automatically recognize sensory effects compared to manual authoring. The results revealed nuanced differences between the modules, showcasing their respective strengths and weaknesses. Analyzing precision, recall, and F1 scores provided a thorough understanding of the overall performance. Notably, the evaluation hinted at potential synergies between the modules, suggesting that their combined use could enhance the effectiveness of mulsemedia content authoring.

Future research could broaden the scope of this assessment by including additional content recognition modules. Some more advanced recognition models, based on transformers [18] can possibly enhance the recognition of sensory effects. This would allow also for a more thorough evaluation of their diverse capabilities of each model and provide a more comprehensive understanding of how automatic extraction impacts authoring efficiency in a wider range of scenarios. A key future work also is to investigate also gender differences impacts on the authoring of sensory effects.

Key considerations for future research include devising strategies for result combination, especially when both modules output similar results. One important venue for work is for the architecture to account for scenarios

where two or more modules exhibit antagonistic outputs. For instance, how would one treat the antagonistic nature of a module recognizing an aroma while another identifies a wind blowing strongly? Should the aroma be considered actively present? These nuanced scenarios present exciting avenues for further investigation, delving into the intricacies of result combination and refining the integration process for a more comprehensive and accurate mulsemedia content authoring experience.

8 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. https://doi.org/10.1109/IJCNN.2018.8489357

[3] Raphael Abreu, Douglas Mattos, Joel Santos, George Ghinea, and Débora C. Muchaluat-Saade. 2023. Semi-automatic mulsemedia authoring analysis from the user's perspective. In *Proceedings of the 14th ACM Conference on Multimedia Systems* (Vancouver, Canada) *(MMSys '23)*. ACM, New York, NY, USA, 249–256. https://doi.org/10.1145/3587819.3590979

[4] Raphael Abreu, Douglas Mattos, Joel dos Santos, Gheorghita Ghinea, and Débora Muchaluat-Saade. 2001. Toward Content-Driven Intelligent Authoring of Mulsemedia Applications. *IEEE Multimedia* 28, 1 (2001), 7–16. https://doi.org/10.1109/MMUL.2020.3011383

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

[6] 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. https://doi.org/10.1111/j.1467-6494.1986.tb00391.x

[7] John Brooke. 1996. SUS: A Quick and Dirty usability scale. *Usability evaluation in industry, Chapter 21* (1996), 189–194. https://www.taylorfrancis.com/chapters/edit/10.1201/9781498710411-35/sus-quick-dirty-usability-scale-john-brooke

[8] Lee Anna Clark and David Watson. 2016. Constructing validity: Basic issues in objective scale development. 7, 3 (2016). https://doi.org/10.1037/1040-3590.7.3.309

[9] 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), 1––35. https://doi.org/10.1145/3233774

[10] 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

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

[12] 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. https://doi.org/10.1287/mnsc.35.8.982

[13] 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* 78 (2019), 19201––19227. https://doi.org/10.1007/s11042-019-7312-2

[14] Douglas Paulo de Mattos and Débora C Muchaluat-Saade. 2018. STEVE: a Hypermedia Authoring Tool based on the Simple Interactive Multimedia Model. In *Proceedings of the ACM Symposium on Document Engineering 2018*. 1–10.

[15] 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. https://doi.org/10.1109/ICNC47757.2020.9049485

[16] Egon Dejonckheere, Febe Demeyer, Birte Geusens, Maarten Piot, Francis Tuerlinckx, Stijn Verdonck, and Merijn Mestdagh. 2022. Assessing the reliability of single-item momentary affective measurements in experience sampling. *Psychological assessment* 34, 12 (2022), 1138.

[17] 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

[18] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2022. Transformers in vision: A survey. *ACM computing surveys (CSUR)* 54, 10s (2022), 1–41.
[19] 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
[20] 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
[21] 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. https://doi.org/10.1109/TVCG.2015.2507591
[22] 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. https://doi.org/10.1145/2993369.2993389
[23] 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. https://doi.org/10.1007/s10209-014-0348-1
[24] 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. https://doi.org/10.1145/3464422
[25] Jum C Nunnally. 1978. An overview of psychological measurement. *Clinical diagnosis of mental disorders: A handbook* (1978), 97–146.
[26] José Pinheiro and Douglas Bates. 2006. *Mixed-effects models in S and S-PLUS*. Springer science & business media.
[27] 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. https://doi.org/10.1109/ICCVW.2017.139
[28] Keith S Taber. 2018. The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in science education* 48 (2018), 1273–1296.
[29] 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. https://doi.org/10.1016/j.image.2012.10.009
[30] Markus Waltl, Christian Timmerer, Benjamin Rainer, and Hermann Hellwagner. 2012. Sensory effect dataset and test setups. In *2012 Fourth International Workshop on Quality of Multimedia Experience - QoMEX*. IEEE, 115–120. https://doi.org/10.1109/QoMEX.2012.6263841
[31] Matthew D. Zeiler and Rob Fergus. 2013. Visualizing and Understanding Convolutional Networks. *CoRR* abs/1311.2901 (2013). http://arxiv.org/abs/1311.2901
[32] 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. https://doi.org/10.1109/CVPR.2018.00775

A NUMBERS RELATED TO VIDEOS IN THE DATASET OF [30]

number	video name	number	video name
1	indy4 3	18	tron legacy official comic con teaser trailer hd
2	babylonad 1	19	babylonad 1 d
3	indy4 2	20	babylonad 1 b
4	centurio	21	babylonad 1 c
5	fringe	22	fastandfurious
6	alien 600 128kbmp3	23	csi
7	fastandfurious 3	24	passwordswordfish 1
8	kick ass debut movie teaser trailer hd	25	pirates 600 128kbmp3 1
9	babylonad trlr 01 1080p dl	26	a chinese ghost story1 xvid
10	indiana jones 4-tlr2 h720p	27	indy4 1
11	babylonad 2	28	babylonad
12	babylonad short	29	iron man 2 trailer official
13	2012 official trailer 4 hd	30	babylonad 3
14	alien-resurrection-teaser-640	31	rambo-tlr2 h720p
15	babylonad shortogg	32	prince of persia the sands of time movie trailer
16	fastandfurious 2	33	babylon ad trailer hd
17	transporter	34	afterlife
		35	fastandfurious 1

Received 31 January 2024; revised 13 July 2024; accepted 19 August 2024