Multimedia Tools and Applications (2019) 78:19201–19227
https://doi.org/10.1007/s11042-019-7312-2

# Crowdsourcing authoring of sensory effects on videos

**Marcello Novaes de Amorim[1] · Estêvão Bissoli Saleme[1] · Fábio Ribeiro de Assis Neto[1] · Celso A. S. Santos[1] · Gheorghita Ghinea[2]**

Received: 24 May 2018 / Revised: 28 November 2018 / Accepted: 31 January 2019 /
Published online: 8 February 2019
© Springer Science+Business Media, LLC, part of Springer Nature 2019

**Abstract**
Human perception is inherently multi-sensorial involving five traditional senses: sight, hearing, touch, taste, and smell. In contrast to traditional multimedia, based on audio and visual stimuli, mulsemedia seek to stimulate all the human senses. One way to produce multisensorial content is authoring videos with sensory effects. These effects are represented as metadata attached to the video content, which are processed and rendered through physical devices into the user's environment. However, creating sensory effects metadata is not a trivial activity because authors have to identify carefully different details in a scene such as the exact point where each effect starts, finishes, and also its presentation features such as intensity, direction, etc. It is a subjective task that requires accurate human perception and time. In this article, we aim at finding out whether a crowdsourcing approach is suitable for authoring coherent sensory effects associated with video content. Our belief is that the combination of a collective common sense to indicate time intervals of sensory effects with an expert fine-tuning is a viable way to generate sensory effects from the point of view of users. To carry out the experiment, we selected three videos from a public mulsemedia dataset, sent them to the crowd through a cascading microtask approach. The results showed that the crowd can indicate intervals in which users agree that there should be insertions of sensory effects, revealing a way of sharing authoring between the author and the crowd.

**Keywords** Mulsemedia content · Sensory effects · MPEG-V metadata · Crowdsourcing · Multimedia authoring · Multimedia annotation

## 1 Introduction

Mulsemedia has been designated as AV (AudioVisual) content augmented with other non-traditional sensory objects such as olfactory, gustatory and haptic [16]. Throughout the last decade, researchers have been exploring this coherent combination of human senses to enhance the Quality of Experience (QoE) of users in mulsemedia applications [1, 13, 25, 28, 35, 38, 40]. Nonetheless, mulsemedia applications face a wide spectrum of research

✉ Marcello Novaes de Amorim
cellonovaes@gmail.com

Extended author information available on the last page of the article.

challenges, many of which are by now traditional in the multimedia community. Of these, we mention rendering, distribution, adaptation, sensory cue integration, building mulsemedia databases, usability, compliance, as well as a lack of rapid prototyping tools [9, 16, 31].

Indeed, many challenges stem from non functional requirements and, in this context, inter-operability is primordial. Towards this end, the MPEG-V standard emerged to provide an architecture and specifications for representing sensory effects [19]. An AV content annotated with MPEG-V Sensory Effects Metadata (SEM) should be able to be reproduced efficiently in many different mulsemedia systems even with actuators from different brands. The process of producing interoperable mulsemedia metadata involves making an MPEG-V compatible XML file that contains entities describing different sensory effect presentation features (beginning and the end of each effect time span, its intensity, fading and so forth). This can be done with the help of an authoring tool, which enhances considerably the overall efficiency of the process [7, 20, 32, 37]. This tool allows authors to abstract the difficulty of writing an XML application, through an intuitive graphical interface whereby they can pick up a movie scene, see what they would feel whilst immersed in the scene, and then, arrange the sensory effects therein. Furthermore, researchers have started developing tools and methods to automatically generate MPEG-V SEM from video content [21, 27]. However, there is much research to be done to produce guidelines for authoring, editing and creating mulsemedia applications. Moreover, the difficulty of capturing many different aspects and turning it into trustworthy sensory effect metadata currently remains an issue. Both are challenges which we address in this paper. Accordingly, in this article, we explore whether a crowdsourcing approach can generate interoperable metadata and boost the process of authoring mulsemedia content.

Despite not being complex in terms of manipulating a tool, the authoring of AV content with sensory effects is a skilled process, at the end of which one would be able to identify and capture different scene details, such as the exact point when an effect starts and finishes, the type of effect to be presented and other similar features. This is a manual and subjective process usually requiring accurate human perception, time as well as the ability to produce a coherent combination of sensory effects and AV content.

In the process of authoring mulsemedia content, individuals are subject to natural bias due to their unique prior experiences and the QoE of users is subject to what they are feeling. Thus, we believe that the combination of a collective common sense to indicate time intervals of sensory effects with an expert fine-tuning is a viable alternative to efficiently generate metadata. Our approach is based on the Galton concepts for Wisdom of Crowds [15]. Following the same principle, Chowdhury et al. [8] built the Know2Look system and obtained satisfactory results by using common sense to determine valid annotations in media content. In our particular case, the end result of using the crowd to author sensory effects is the insertion of *Wind* effects and *Vibration* in the intervals in which the crowd agrees that it makes sense to insert them.

Accordingly, in the study reported here, we selected three videos from a public mulsemedia dataset [39]. Next, we sent these to workers, recruited on the MicroWorkers[1] platform, through a cascading microtask approach in which quality will be managed during the different execution stages of this process. Finally, we compared the results of our crowdsourcing approach with the annotations of the public mulsemedia dataset. Results revealed that the

[1]MicroWorkers platform available at http://ttv.microworkers.com.

sensory effects generated by the crowd can be used to validate author insertions as well as to supplement authorship with new sensory effects.

The remainder of the text is organized as follows. Section 2 reviews other studies related to the authoring of mulsemedia content and crowdsourcing approaches for multimedia authoring and annotation. Section 3 presents the workflow that guides our crowdsourcing approach for authoring mulsemedia content. Section 4 describes the tools that support our approach. Section 5 presents the evaluation of the work. Section 6 discusses the results. Finally, Section 7 concludes the article and highlights future work.

### 1.1 Scope of this work

It is important to clarify from the outset that this article presents a crowdsourcing approach for sensory effects video authoring. Strategies or new features for its annotation are out of scope here. In our approach, workers make trivial annotations on videos, so the tools used to collect them are simple HTML documents that contain players and additional controls to indicate instances and intervals. Subsequently, the collected annotations are filtered, grouped, and aggregated to produce results that represent the common sense from the crowd. There are models and definitions that aim to standardize the production of annotated media, such as the canonical process presented by Hardman et al. [18] for semantically annotated media production. However, the model presented in this work is related to the cascade microtasking process of crowdsourcing, not to the annotation process itself. The annotation collected from the crowd is straightforward and used as an input to the aggregation method, which in turn, processes the contributions and generates the final results.

Likewise, it is not within the scope of this article to propose a presentation model for mulsemedia videos, for instance similar to that presented by Sadallah [29] for hypervideo on the Web. Instead, the authoring process provides interoperable and shareable mulsemedia data in MPEG-V, which allows results accessed from compatible systems.

## 2 Related work

Mulsemedia authoring tools have been developed for almost a decade. SEVino [37], SMURF [20], RoSE Studio [7], and Real 4D studio [32] are examples of tools that support authors in adding sensory effects, usually represented as MPEG-V metadata, to AV contents. Players compatible with MPEG-V such as Sensorama [6], SEMP [37], Sensible Media Simulator [20], Multimedia Multisensorial 4D Platform [4], and PlaySEM [30] are able to reproduce this kind of authored content. Thus, all of these tools shape an ecosystem for delivering and rendering mulsemedia content.

The work of Kim et al. [21], and more recently the work of Oh and Huh [27], represent an attempt to automatically generate interoperable mulsemedia metadata. The authors argue that this method can speed up the authoring process, helping the industrial deployment of mulsemedia services. Kim et al. [21] proposed a method and an authoring tool to extract temperature effect information using the color temperatures of video scenes and generate MPEG-V SEM. The authors found that the automatically generated temperature effects enhanced the level of satisfaction of the users through a subjective evaluation. However, its limitation to generate only one effect relies on the recurrence of manual tools to add other effects. Oh and Huh [27] proposed a similar approach to automatically generating MPEG-V SEM based on the motion of an object included in a media. Akin to the approach of Kim et al. [21], it is limited to the temperature effect information, automatically extracted from

the color temperatures of video scenes. Furthermore, the authors did not show the results of the method, which makes it difficult to evaluate its efficiency.

Interoperability is an important issue to be addressed in systems for authoring and annotating media. This problem is often addressed in works in this area, such as Sadallah et al. [29], which presents a high-level component-based model for specifying annotation-driven systems focused on hypervideo. In this context, it is interesting to adopt models and standards that promote the interoperability of generated metadata, such as the canonical process proposed by Hardman et al. [18] for the production of semantically annotated media. The metadata interoperability allows its use in different applications, including by automatic methods, as can be observed in Ballan et al. [3] that has surveyed works focused on the detection and recognition of events for semantic annotation of videos.

Over the past years, many studies involving multimedia content processing applied crowdsourcing approaches, such as for generating video annotations [11, 23], image descriptions [14, 33], real-time captioning [22], text annotations [12, 42], and audio annotations [24, 34]. Recently, See and Chat [5] demonstrated how to automatically generate annotations in user comments on images published on Flickr[2].

With regard to the generation of complex video metadata using crowdsourcing methods, the work of Cross et al. [10] presented the VidWiki system, which is a complex application to improve video lessons. This system requires that workers edit video scenes by adding annotations, including complex annotations such as LaTeX equations. It also requires that the recruited workers have previous knowledge about LaTeX. Another crowdsourcing complex video annotation system was proposed by Gottlieb et al. [17], which achieved geo-location annotations for random videos from the web by requiring the workers to perform searches on the internet, use encyclopedias and provide annotations in the specific format for GPS coordinates.

In relation to the crowdsourcing processes for multimedia authoring, also noteworthy is the work of Amorim et al. [2], which used a process composed of cascaded microtasks to generate interactive multimedia presentations from plain videos. In that work the audience was responsible for identifying the points of interest to be enriched, as well as making available, selecting and positioning the media content in the video to generate the multimedia composition.

However, as far as authoring of mulsemedia content based on a crowdsourcing approach is concerned, we did not find related studies dealing with the matter.

## 3 A crowdsourcing approach for authoring of sensory effects

Our process for mulsemedia content authoring is composed of a sequence of microtasks, thus enabling to use the workforce of a plethora of unskilled workers towards mulsemedia authoring. In fact, this process can be viewed as a generic solution that is tailored to different types of crowdsourcing annotation projects. Figure 1 presents the three phases - *Preparation*, *Execution*, and *Conclusion* - of this generic process.

The *Preparation Phase* deals with the definition of AV content to be annotated, the source of the contributions for each task, the monetary resources to pay workers, and the tools used by them for performing the tasks.

[2]Flickr available at https://www.flickr.com.

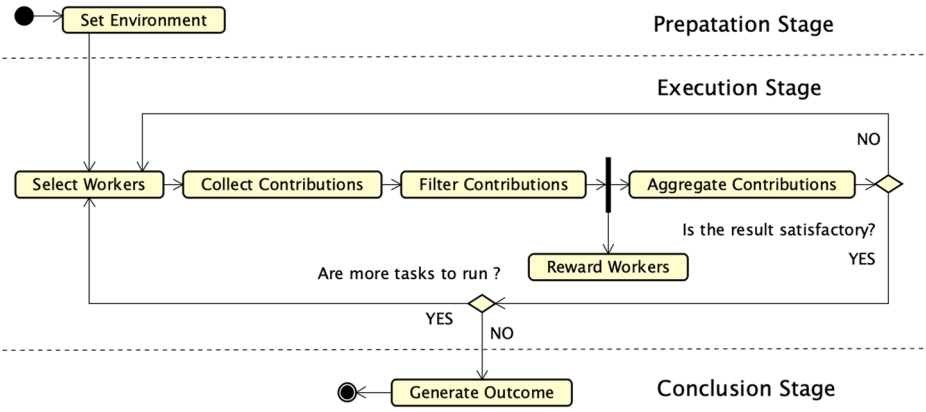



**Fig. 1** Generic crowdsourcing process template's workflow

The *Execution Phase* deals with the part of the workflow for content annotation, which is generically described in the form of a complete algorithm that controls task flows within time, cost and quality constraints, to reach a desired end result. In our approach for mulsemedia authoring, crowdsourcing tasks are performed in cascade, namely a sequence of several similar stages with each stage processing the output from the previous stage. In addition, all task executions are associated to the same sequence of steps: *Selecting Workers*, *Collecting Contributions*, *Filtering Contributions*, *Reward Workers*, and *Aggregating the Filtered Contributions*. The results generated by the aggregation method for a task provides the input for the next one. Once the output of the aggregation method is satisfactory, the process advances to the next stage in the cascade for further processing. Otherwise, the current task in the workflow must be restarted to select new workers, collect new individual results and update the aggregated task results. *Reward Workers* are located after the *Filter Contributions* to cover cases where the *Owner* has decided to pay only for valid contributions. Thus, if the process requires another task, simply follow that model, defining the annotation tool, the filters and the aggregation method, and insert it into the process.

In the *Conclusion Phase*, the end result is produced and evaluated using a specific method defined in the project.

The generic process described in this section can be used to create crowdsourcing workflows to coordinate the crowd through a sequence of tasks, managing their dependencies, and bringing together intermediate results produced by the workers as in the case of Fig. 2.

In fact, a workflow represents an effortless way to understand the whole process since, from hiring workers to processing itself, the results are provided by the crowd. In addition, as stated by Assis Neto and Santos [26], crowdsourcing workflows are context-oriented and they should establish not only how the process activities will be performed by the crowd, but also how the quality will be managed through the different execution stages of the crowdsourcing process.

### 3.1 The workflow for mulsemedia authoring

The workflow presented in Fig. 2 represents our crowdsourcing process for mulsemedia authoring based on AV content annotation with MPEG-V SEM. This workflow consists of

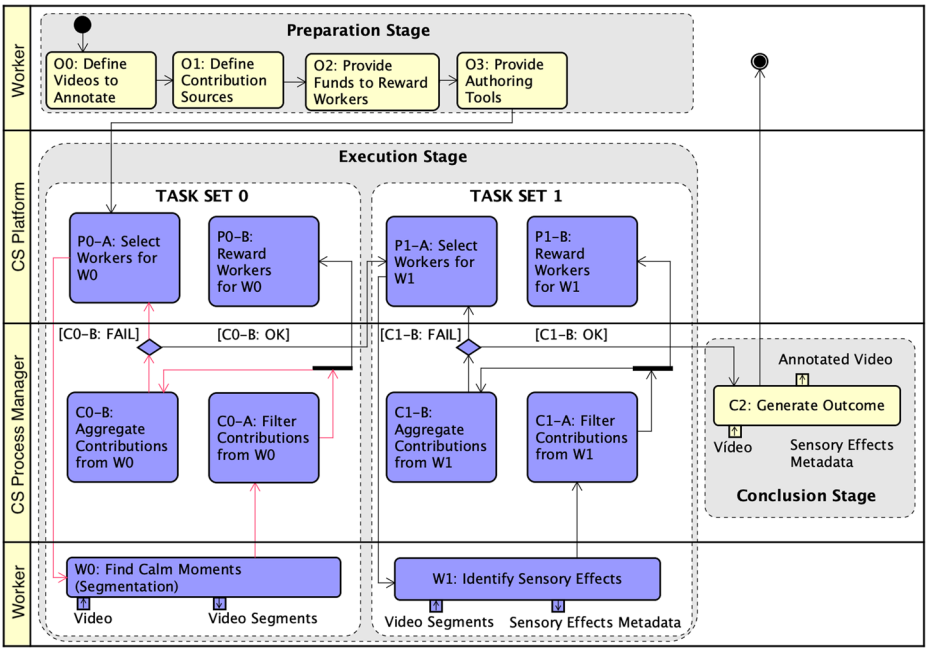



**Fig. 2** Crowdsourcing process's workflow for mulsemedia content authoring

a set of tasks performed by actors performing four roles: *Owner*, *Crowdsourcing Platform* (*CS Platform*), *Crowdsourcing Process Manager* (*CSPM*), and *Worker*.

The process begins with the *Preparation Stage* (see Fig. 1) in which the *Owner* sets the environment to start the process. In our view, the *Owner* is someone who works with a crowdsourcing management team, composed of experts in the field, and responsible for specifying the technical requirements as well as qualified personnel to create the tasks.

At the beginning of this stage, the owner performs the activity ***O0: Define Videos to Annotate***, registering the videos to be annotated. Next, he/she must then perform the activity ***O1: Define Contribution Sources***, in which he/she chooses whether to use a commercial crowdsourcing platform or other mean to reach workers to collect their contributions. When choosing to use a commercial crowdsourcing platform, it is necessary to deposit funds to reward workers, this is done in activity ***O2: Provide Funds to Reward Workers***. Completion of the *Preparation Stage* is reached when the activity ***O3: Provide Annotation Tools*** is concluded and the campaigns are created in the *CS Platform*. In our case, the campaigns correspond to the crowd tasks ***W0: Find Calm Moments*** and ***W1: Identify Sensory Effects***.

The *Crowdsourcing Platform* is the source of contributions, and is responsible for the activities ***P0-A and P1-A: Select Workers for W0 and W1***, and ***P0-B and P1-B: Monetary Reward Payment for W0 and W1***.

The *Crowdsourcing Process Manager*(*CSPM*) represents a person, or a team, responsible for monitoring the process and analyzing the current state of contributions to decide when a task should be closed, as well as initiating the generation of results for each task, together with stopping and starting them.

The *CSPM* is responsible for activities that produce partial results and compile into the outcome.

**Table 1** Responsibilities and activities for each role

| Role | Responsibilities | Activities |
|---|---|---|
| Owner | Setting up the environment; Provide funds to pay Workers. | O0, O1, O2 and O3 |
| Crowdsourcing Platform (CS Platform) | Recruiting Workers; Intermediate payments. | P0-A, P0-B, P1-A and P1-B |
| Crowdsourcing Process Manager (CSPM) | Manage the process workflow; Initiate Filtering and Aggregation; Control transitions between tasks. | C0-A, C0-B, C1-A, C1-B and C2 |
| Worker | Execute the annotation tasks. | W0 and W1 |

In the activities, ***C0-A and C1-A: Filter Contributions from W0 and W1*** reliability filters are applied over the collected annotations from the crowd, so activities ***C0-B and C1-B: Aggregate Filtered Contributions from W0 and W1*** can then process reliable contributions to construct the results. Finally, after all the partial results are made, *CSPM* executes activity ***C2: Generate Outcome*** to export the annotated video to the desired format.

The *Workers* are responsible for providing the information required to produce the outcome. They are responsible for performing the annotation tasks by executing the activities ***W0: Find Calm Moments*** and ***W1: Identify Sensory Effects***.

Responsibilities and activities for each role are summarized in Table 1.

## 4 CrowdMuse: Crowdsourcing mulsemedia authoring system

We developed the *CrowdMuse* system to support our crowdsourcing approach. One of the most important characteristics of the system is its capacity of distributing tasks to workers from various sources, such as commercial crowdsourcing platforms, internal teams, and social networks.

*CrowdMuse* follows a component-based approach to manage the complexity of a mulsemedia annotation problem by breaking it down into smaller and physical manageable modules. The modules *Server*, *Client*, and *Persistence* in Fig. 3 are the units of implementation of the *CrowdMuse* system and are assigned areas of functional responsibility. In addition, the work interfaces in the system were constructed as simple HTML-5 documents, which just render information coming from the *Server Module* and send back contributions.

Another advantage of the *CrowdMuse* architecture is that even when using commercial crowdsourcing platforms, the collected data is stored only in the system database. Moreover, this system is responsible for controlling the execution flow of the tasks, managing the items that must be annotated, generating the jobs that must be executed and then distributing these jobs among workers.

### 4.1 The persistence module

In the *Persistence Module*, the *Crowd Knowledge Base* component sends to the server module the information required to render the job requests to workers and the content needed to present the result to the users. Likewise, contributions produced by workers were sent directly from the collector to the *Crowd Knowledge Base* component, in which they were

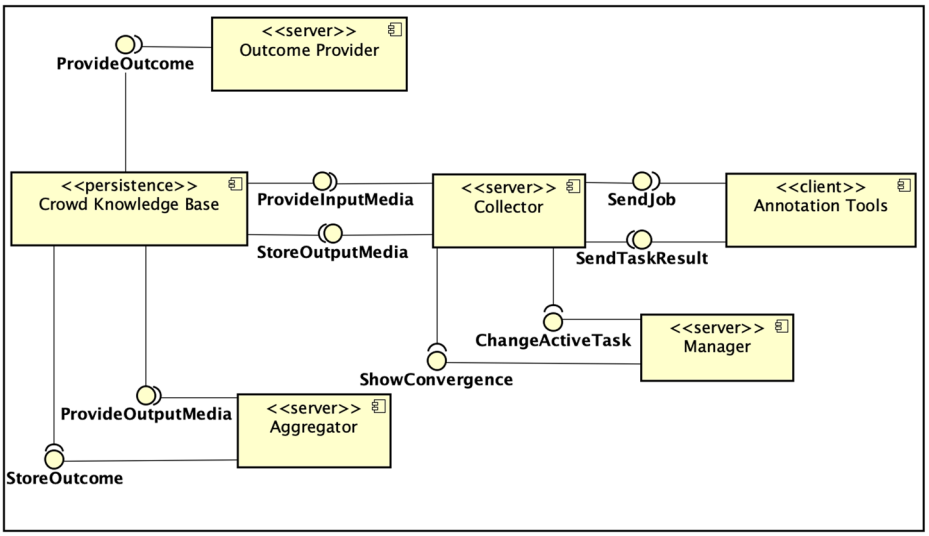



**Fig. 3** CrowdMuse system components and communication interfaces

stored directly in the database without having to go through the external crowdsourcing environment.

The *Aggregator* component also communicates directly with the persistence. The *Aggregator* retrieves the collected contributions of a task set and, after the aggregation process, sends the result to be stored in the database, to be used as input to the next task, again maintaining data privacy because it does not need to be stored in an external environment.

### 4.2 The server module

The server module is responsible for distributing the jobs, managing contributions, and controlling the active task in order to execute the process workflow. This module is composed of four components: *Manager*, *Collector*, *Aggregator*, and *Player Provider*.

- **Manager** is the module responsible for controlling the enrichment process. It is related to the task-to-task transition tool, as well as the tool to monitor the current state of tasks and trigger aggregation methods. Management functionality is accessible through the management interface.
- **Collector** provides the annotation tool with information about the item to be annotated, therefore, it renders the job's interface used by the worker to perform the task. Also, this component is responsible for gathering the information provided by the worker after the execution of the task and sends them to the persistence.
- **Aggregator** applies reliability filters over the worker's contributions and processes the valid annotations in order to generate the result for each task. The aggregation methods are based on convergence analysis to produce collective results.
- **Player Provider** delivers the process outcome that consists of mulsemedia annotated videos. This outcome can be exported and visualized in players able to reproduce these effects, such as SEVino [37] tool.

### 4.3 The client module

The *Client Module* manages the communication interfaces involving workers and other users. This module presents templates that generate visualization data according to a description. For each task of the process, the client must render a specific annotation tool for the worker. Thus, it is possible to keep the server accessible through the *Collector* and *Player* components, and therefore templates can be stored from anywhere. This allows contributions to be collected from different workers at the same time. Moreover, a model is selected and the Persistence module is queried to obtain the necessary data to render the desired interface according to the desired data visualization.

### 4.4 The public interfaces

The communication between the modules of the *CrowdMuse* system occurs through the public interfaces represented in Fig. 3 and detailed as follows:

- **Change Active Task:** The *Owner* sends a request to the Server to set the currently active task.
- **Show Convergence:** The *Manager* displays to the owner the current convergence state for the active task.
- **Provide Media Input:** To generate each job, the *Collector* receives an entry from the *Persistence* component.
- **Send Job:** The *Collector* sends a job to a worker who sees the task through the *Client* and executes it.
- **Send Task Result:** The *Client* sends to the *Collector* the annotation made by the Worker.
- **Store Media Input:** The *Collector* sends workers contributions to the *Persistence*, that stores it in the *Crowd Knowledge Base*.
- **Provide Output Media:** The *Persistence* send to the *Aggregator* all the contributions collected related to a task.
- **Store Outcome:** The *Aggregator* stores the entries received from the aggregation process. The generated outcome can be provided as input to the next task.
- **Provide Outcome:** The *Persistence* module provides the outcome of the crowdsourcing project (i.e. a set of MPEG-V SEM) to be rendered with the video content.

### 4.5 Considerations

The CrowdMuse system can be freely used and modified to serve different crowdsourcing applications with a focus on authoring of mulsemedia and other kinds of multimedia content. In the next section, we will present a case study demonstrating the use of CrowdMuse for crowdsourcing authorship of mulsemedia content according to our approach.

## 5 A case study on crowdsourcing authoring of mulsemedia content

A case study concerning the crowdsourcing authoring approach proposed her was carried out using three from a public mulsemedia dataset [39], referenced in the paper as *Babylon A.D.*, *Formula 1*, and *Earth*. As stated by Timmerer et al. [35], these three videos have obtained the highest MOS in their QoE experiments which were performed over this same

**Table 2** AV content annotated with sensory effects used in the experiment

| Video | Resolution (WxH@fps) | Bit-rate (Mbit/s) | Duration (s) | Wind | Vibration |
|---|---|---|---|---|---|
| Babylon A.D. | 1280x544@24 | 6.81 | 118.42 | 10 | 8 |
| Earth | 1280x720@25 | 6.90 | 66.00 | 7 | 1 |
| Formula 1 | 1280x720@25 | 5.40 | 116.2 | 11 | 4 |

dataset and that is the main reason for this choice. Despite not having enough evidence, we assumed that workers could perceive sensory effects easier in videos with high MOS than in random videos, and thereby give a clearcut contribution.

The reference mulsemedia dataset also contains information about the intensity of some sensory effects. However, because there was no homogeneity between the audio and video equipment used by the workers, it was decided not to request that they observe the intensity of the effects, only the intervals at which they should be inserted. In addition, we noticed that in the dataset of Waltl et al. [39] there are annotations of effect with very subtle intensity, and we chose not to consider them for this experiment, believing that a more specialized work would be needed to accurately record the subtle effects.

We decided to collect the same kind of effects associated by of Waltl et al. [39] to the selected videos, that is *Wind* and *Vibration* effects. The metadata associated with lighting information, also annotated in the videos, will be not considered in our experiment once it is set to be auto-extracted according to the brightness and color information of the video frames.

Table 2 presents the main characteristics of the three videos annotated with sensory effects used in our evaluation.

To conclude this section, we come to the first question posed in the introduction of the paper: Is the crowd capable of producing a coherent and cohesive set of sensory effects related to the AV content processed by each worker individually? Other questions have to do with the effort and quality of the content produced in a crowdsourcing process.

### 5.1 Setting the environment

According to the workflow of Fig. 2, the *Owner* should perform four activities to set up the environment before beginning to collect contributions from the crowd. The first activity is ***O1: Set videos to annotate*** and consists of selecting the videos that should be annotated. These videos should be uploaded to a video repository and made available to workers. In the activity ***O2: Define Contribution Sources***, the *Owner* chooses if the contributions will be collected from a contracted crowd using a commercial platform or the workers will be volunteers or members of internal groups.

The Owner may also need to create a campaign on the crowdsourcing platform (Microworkers, in our case) for each task in the process workflow. To create a campaign, the *Owner* must perform the activity ***O3: Provide Funds to Reward Workers*** to ensure the means to pay workers, and the activity ***O4: Provide Annotation Tools*** in which the annotation tool that will be used to perform the task is sent to the crowdsourcing platform.

### 5.2 Crowd definition

As mentioned before, *Microworkers* performs the role of *CS Platform* in the project workflow of Fig. 2. Thus, this crowdsourcing plataform is responsible for recruiting and paying the workers, although *CrowdMuse* is compatible with other commercial platforms, such as Amazon's Mechanical Turk (AMT)[3] and CrowdFlower[4].

*Microworkers* proposes different models for a crowdsourcing project. Initially, one can choose between starting a basic campaign or using contracted groups. In a basic campaign, all registered workers in the platform can see the task in the job menu and work on it. A campaign that uses hired groups allows the *Owner* to select the crowd by choosing groups of workers with the desired profile. In addition, it is possible to create lists with workers who have made good contributions before, so they can be recruited to work on other tasks.

One of the characteristics of our approach is that it uses very simple tasks and unskilled workers. The tasks were launched as campaigns that used contracted groups, to increase the chance of the workers who contributed to a task also participating in others. A group of moderate size was chosen so the contributions were made quickly. The group chosen is identified as *Data Services* in *Microworkers* platform, with 1285 potential workers to accept the jobs. Some groups were composed of workers who only accepted tasks that offered slightly larger payment, but considering the chosen group, it was feasible to offer a payment of 0.05 USD per task.

### 5.3 Method

As shown in the workflow of Fig. 2 the crowdsourcing process for authoring mulsemedia content is based on two microtasks executed in cascade. Each microtask is executed as a complete task-set construction composed of three sequential main activities: (i) contributions collection, (ii) filtering, and (iii) aggregation. Hence, the individual contributions are collected from each worker through a specific tool required for executing the assigned task. In the sequence, the contributions are filtered and clustered using the aggregation function to extract the results of the microtask execution.

The next two subsections will describe the two tasks (***Find Calm Moments*** and ***Identify Sensory Effects***) that made up the crowdsourcing process for mulsemedia authoring.

### 5.4 The first crowd task: Find calm moments (Segmentation)

The objective of the first task was to segment the video in such a way that the sensory effects were not fragmented by more than one segment, that is, the effects contained in a segment should be completely contained in it. In this task, one of the three selected videos was displayed to the worker who should indicate the instants that he/she thought there is no *Wind* or *Vibration*, pretending that he/she was immersed in the environment of the movie.

#### 5.4.1 Contributions collection

We developed the tool shown in Fig. 4 to support the first microtask. In this tool, the video to be processed is on the bottom of the window, whilst the buttons used by the worker to

[3] AMT - https://www.mturk.com.

[4] CrowdFlower - https://crowdflower.com.

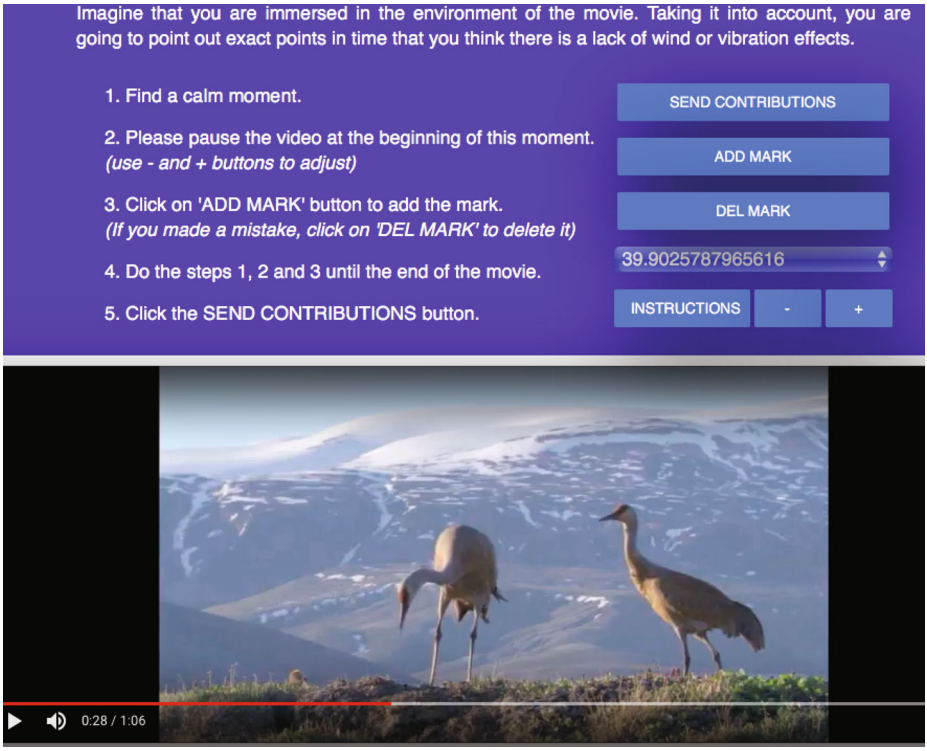



**Fig. 4** Tool for identifying calm moments in a video

determine the calm moments on the video as well as the task instructions are on the top. As discussed, the instants pointed out by the crowd are used later for content segmentation.

In each contribution, a worker could supply as many time marks as he/she wanted, each mark representing the initial instant of a calm moment in the video. In this task execution, 23 contributions were obtained that provided 113 time-points for the *Babylon A.D.* video, 17 contributions that provided 213 points for the *Formula 1* video, and 21 contributions that provided 108 points for the *Earth* video.

### 5.4.2 Filtering contributions

The collected contributions were filtered according to (i) the number of time-points provided and (ii) the proximity of these time-points.

With each task, workers could annotate multiple marks in the video segment. Thus, the first quality criterion was to calculate the average number of marks received by each segment and to discard the contributions containing differing amounts of marks. Contributions with less than 50% or more than 200% of the average number of marks were discarded. In addition, very close marks of the same worker contribution were discarded. It was established that the segments should be at least 0.5 s in length. Thus, when a worker provided two separate marks for less than 0.5 s, the first annotation was ignored as it was assumed that an update had occurred and the worker forgot to delete the previous annotation.

**Table 3** Contributions and results for the first task

| Video | Contributions | Calm moments | Filtered | Segments |
|---|---|---|---|---|
| Babylon A.D. | 23 | 113 | 68 | 11 |
| Earth | 21 | 108 | 59 | 12 |
| Formula 1 | 17 | 213 | 179 | 15 |
| Total | **61** | **434** | **306** | **38** |

Summarizing, a total of 68 over 113 time marks remained after the filtration stage for *Babylon A.D.* video. For *Earth* video, 59 over 108 time marks and for *Formula 1* video, 179 over 213 time marks were delivered to the aggregation stage.

### 5.4.3 Aggregation

The aggregation process for the first microtask is based on the grouping of the contributions so that each group contains suggestions from the crowd regarding the same calm instant in AV content. Because marks are point values that represent instants of the video, the algorithm used is based on neighborhood grouping. This strategy assumes that the distance between two marks referring to the same calm moment tend to be closer than the marks for consecutive calm moments.

In our aggregation strategy for this task, the marks were sorted in ascending order, and a $\Delta$ value was calculated that represents the average distance between the consecutive marks provided. This $\Delta$ was then used as the threshold for the grouping. When the distance between one time-point contribution and the next one is greater than $\Delta$, a new group is started.

At the end of this stage, each group obtained represents the initial instant of a calm moment for which the crowd agreed to exist. Therefore, time-points that do not fit into any group were discarded.

The video segments are determined using the calm moments defined by the crowd. Each video segment to be annotated with sensory effects is associated to the time interval between two calm moments. With this strategy, we aimed to obtain segments that contain *Wind* and *Vibration* effects without being fragmented by more than one video segment. In this way, a total of 11, 12, and 15 segments were obtained for the *Babylon A.D.*, *Earth*, and *Formula 1* video, respectively (see Table 3).

Finally, the segments to be annotated with sensory effects thus determined were used as input for the second microtask in our authoring approach.

Table 3 shows that, out of a total of 61 contributions, 434 suggestions of calm moments were observed by the crowd. After filtering, that produces 306 instants, whilst 38 segments were obtained after running the aggregation stage.

## 5.5 The second crowd task: identify sensory effects

The second task asks workers to provide subjective information, aimed at obtaining ranges in which *Wind* or *Vibration* effects should be pointed out. We re-enforced the workers to provide the maximum time ranges they could and be trustful in an attempt to receive more reliable contributions in this task. We did not ask for intensity because we believe it is a fine-tuning task that requires expert skills such as fade-in and fade-out. However, we advised the workers to create a new range of the same effect if they realized a change in intensity. The

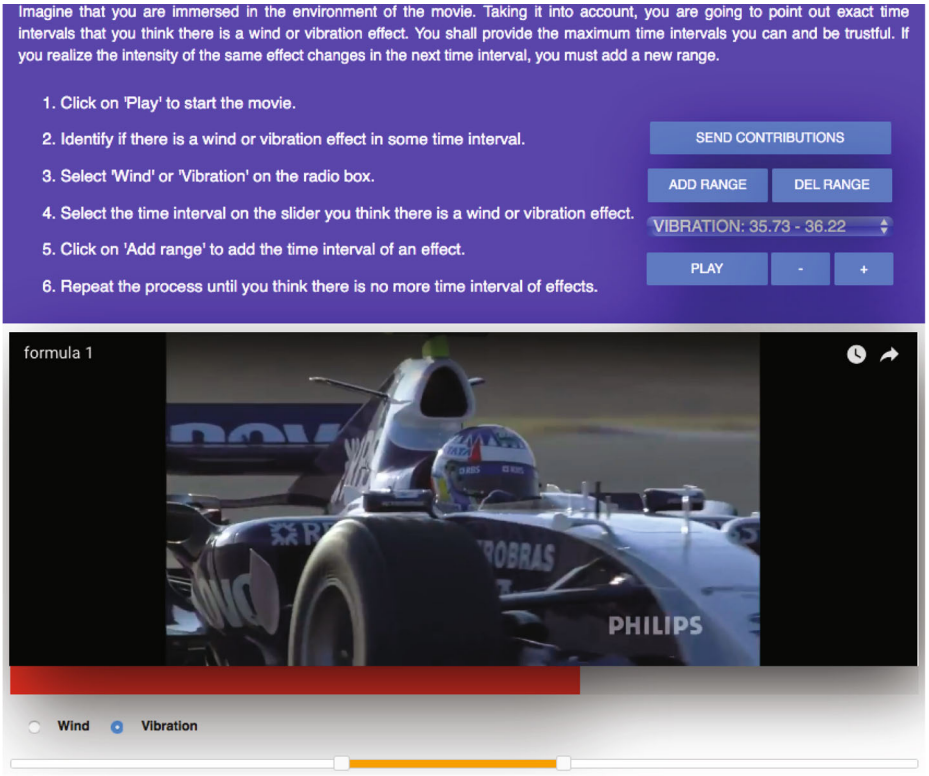



**Fig. 5** Tool for identifying sensory effects

input of this second task was the set of 38 video segments produced in the previous task, as detailed in Table 3. The segments resulting from the aggregation of the contributions collected in the first microtask were delimited by moments of total absence of effects, so that it is possible that in these second microtasks more than one insertion of effect within each segment is identified. This occurs in cases where two inserts of high intensity effects are separated by a low intensity sensory effect, without ceasing completely the effect. In this way, it is possible and acceptable that the number of sensory effects identified in the videos at the end of this task is greater than the number of segments received as input.

### 5.5.1 Contributions collection

We implemented the tool depicted in Fig. 5 to support the collection activity related to the second microtask. The look-and-feel is very similar to the tool presented in Fig. 4. The instructions followed by the workers and buttons for correctly performing the task are presented, at the top, and the analyzed video, at the bottom.

The second microtask was executed in two stages: the first comprised 60 jobs, whilst in the second another 90 were executed, totaling 150 jobs. In each job, a worker annotated one or more time spans in which he/she believed that the effects of *Wind* or *Vibration* should be inserted. The output was a list of time ranges of *Wind* and *Vibration* effects. The 150 contributions provided 166 ranges for insertion of *Wind* effects and 146 of *Vibration* effects.

**Table 4** Contributions and results for the second task

| | Wind | | | Vibration | | |
|---|---|---|---|---|---|---|
| Video | Ranges | Filtered | Converged | Ranges | Filtered | Converged |
| Babylon A.D. | 28 | 25 | 8 | 48 | 46 | 10 |
| Earth | 65 | 57 | 11 | 60 | 46 | 5 |
| Formula 1 | 53 | 49 | 10 | 58 | 55 | 16 |
| Total | **146** | **131** | **29** | **166** | **147** | **31** |

### 5.5.2 Filtering

In an attempt to eliminate malicious and inconsistent contributions, two filtering criteria were used: (i) amount of ranges provided in the contribution, and (ii) existence of overlap between the ranges provided in the contribution.

To meet the first criterion, the average number of ranges in a same contribution for each segment was calculated, and contributions that received less than 50% or more than 200% of that amount were eliminated. The second criterion evaluated the existence of overlap between ranges provided by the same worker for a given effect, in which case it was assumed that the range was updated but the worker forgot to delete the first annotation, so only the most recent ranges of each overlap was maintained .

At the end of the filtering process, 131 of the 146 *Wind* effects identifications and 147 of the 166 *Vibration* ones remained. For the video *Babylon A.D.*, there were 25 identifications of the *Wind* and 46 of the *Vibration* effects; for *Earth* video, 57 identifications of the *Wind* and 46 of the *Vibration* effects and, finally, for *Formula 1* video, 49 identifications of the *Wind* and 55 of the *Vibration* effects.

### 5.5.3 Aggregation

The aggregation of the contributions collected in this second microtask is based on grouping the contributions so that each group contains suggestions from the crowd regarding the same time range for adding a sensory effect.

Firstly, the intervals were divided by video and subdivided by type of effect, *Wind* or *Vibration*. Then each division is ordered in relation to the initial benefit of the range. Finally, a grouping of the intervals is performed so that each group is composed of overlapping ranges. Non-overlapping ranges were discarded. Each of these groups represents a crowd-agreed range for insertion of sensory effects. In this way, the aggregation function was applied to each group to determine the convergent ranges.

The aggregation function determines the maximum degree of overlap between contributions. Then, this maximum degree is used as a boundary to adjust the upper and lower limits of each range in order to delimit the wider region with degree of overlap greater than half of the maximum.

Table 4 shows the numbers of ranges for *Wind* and *Vibration* effects provided by the crowd for each video and its processing. The 150 contributions collected from the workers provided 312 notes of sensory effects, of which 146 were *Wind* and 166 were *Vibration*. After filtering, only 278 of the initially proposed 312 effects were carried forward to the aggregation stage, which in turn delivered 29 *Wind* and 31 *Vibration* effects to be annotated in the selected videos.

### 5.6 Conclusion stage

The conclusion stage aims to generate the crowdsourcing project outcome. At this point, the internal representation of the sensory effect metadata for each selected video is already created. To promote interoperability, the final result of the process is represented in keeping with the MPEG-V format, so that the results generated through this work can be refined with the help of tools like SEVino [37] and Real 4D studio [32], and rendered by mulsemedia players such as PlaySEM [31] and SEMP [37]. The results could also be represented in the format EAF (ELAN Annotation Format), compatible with the ELAN [5] video annotation system which, in addition to allowing the result to be displayed, can also export it to other formats.

## 6 Results and discussion

In order to analyse the results of our study, we made comparisons of the content produced by our crowdsourcing approach, using the three videos selected from the database of Waltl et al. [39] (i.e. *Babylon A.D.*, *Earth* and *Formula 1*), with the annotations for the same three videos, produced by a specialized team responsible for populating this public database. Although the public database contains information about the corresponding sensory effects and their attributes, as intensity, we decided that crowd members should only determine the corresponding, *Wind* or *Vibration*, sensory effect to each video scene, irrespective of the intensity of the annotated effect.

It is noteworthy that the comparison between the effects identified by the crowd and those identified by the author, rather than measuring the similarity between the results, aims to understand how they complement each other.

### 6.1 Babylon A.D.

The video is a commercial trailer for an action movie that features mainly gunshots and explosions. The workers contributed 38 times to it, indicating 28 *Wind* and 48 *Vibration* effects to the video. After running the task aggregation method, 8 time intervals containing *Wind* and 10 time intervals containing *Vibration* remained. The most noticeable events on this video corresponding to gunshots and explosions were identified by the crowd, including some which hadn't previously been annotated in the reference dataset. Moreover although the two most perceptible explosion events in the reference dataset were not obtained using the aggregation method, these events received contributions from the crowd participants. Tables 5 and 6 show the effects of *Wind* and *Vibration* obtained from the crowd compared to the annotations of the reference dataset [39].

While analyzing the content of the *Babylon A.D.* trailer, it was possible to identify why some effects were present in the dataset and not perceived in the same way by the crowd. For instance, the beginning of the video presents an object (a satellite) moving through space. Although the dataset metadata had a *Wind* effect annotated, the workers did not indicate that, probably because they considered that there is no air flow in space. This occurrence demonstrates how the author tends to use the sensory effects to convey his artistic vision of the scene. Furthermore, the crowd tended to associate effects to high motion scenes, such

[5] ELAN available at https://tla.mpi.nl/tools/tla-tools/elan.

**Table 5** Babylon A.D. - Vibration

| Author | | Crowd | |
|---|---|---|---|
| start | end | start | end |
| | | 8.09 | 10.99 |
| 26.7 | 30.7 | 29.04 | 30.94 |
| | | 32.85 | 33.55 |
| | | 39.28 | 39.98 |
| | | 47.02 | 47.32 |
| 49.2 | 49.6 | 48.60 | 57.90 |
| 61.6 | 63.6 | 61.30 | 67.00 |
| | | 69.40 | 73.00 |
| 74.7 | 74.9 | | |
| 75.5 | 75.8 | | |
| 78.0 | 78.5 | | |
| 89.6 | 99.2 | 98.04 | 98.24 |
| 99.2 | 109.2 | | |
| | | 112.00 | 117.20 |

as those depicting explosions, whereas the reference dataset mainly associated effects to scenes with low motion.

By analyzing the annotated video it is possible to verify that the most evident events were identified by the author and the crowd. Gun shots and explosions of lesser intensity were noticed only by the crowd, while the author's explicit annotations refer to subtle events. In this way, the effects identified by the author and by the crowd are complementary, covering both the workers' expectations and the subtleties intended by the author.

**Table 6** Babylon A.D. - Wind

| Author | | Crowd | |
|---|---|---|---|
| start | end | start | end |
| 10.4 | 11.8 | | |
| 12.4 | 33.5 | 29.31 | 33.41 |
| | | 36.00 | 39.20 |
| 40.0 | 44.3 | 42.07 | 45.97 |
| 45.9 | 49.2 | 47.00 | 47.10 |
| 53.8 | 55.0 | 54.09 | 57.99 |
| 63.1 | 63.6 | | |
| 73.8 | 74.6 | | |
| 75.5 | 75.9 | | |
| | | 92.47 | 93.07 |
| 97.2 | 99.0 | | |
| 102.6 | 109.2 | 105.70 | 106.90 |
| | | 113.08 | 114.88 |

**Table 7** Earth - Vibration

| Author | | Crowd | |
|---|---|---|---|
| start | end | start | end |
| | | 3.40 | 6.30 |
| 18.3 | 18.7 | | |
| | | 21.0 | 27.70 |
| | | 37.40 | 39.20 |
| | | 52.90 | 53.70 |
| | | 57.25 | 57.75 |

### 6.2 Earth

The *Earth* trailer had 63 contributions, resulting in a total of 65 *Wind* effects and 60 *Vibration* effects. After running our aggregation method, 11 time intervals containing *Wind* and 5 time intervals containing *Vibration* remained. The workers noticed more *Vibration* effects in this video than the public dataset. Taking into account the analysis of the *Earth* scenes, we concluded that the workers perceived *Vibration* in scenes with stronger movements, such as when an animal jumps or in a herd of animals running. Moreover, when they heard a very loud noise in scene transitions, they pointed *Vibrations*. Tables 7 and 8 show the effects of *Wind* and *Vibration* obtained from the contributions of the crowd, compared to the annotations present in the reference dataset [39] with an intensity greater than or equal to 50%.

Regarding the *Wind* effect, the workers associated it with fast movements, e.g. scenes with cloud movements and a herd of running animals. They also noticed *Wind* in a scene where a quick presentation of slides with animal images was displayed.

It was evident in this video that the crowd complemented the effects indicated by the author, adding *Wind* effects to scenes that featured fast movements and *Vibration* effects for scenes with strong movements.

### 6.3 Formula 1

This AV content is an advertisement for *Formula 1* racing, in which there are several scenes of racing cars as well as scenes of pit stops and pilots walking. The workers made 49 contributions, resulting in a total of 53 *Wind* effects and 58 *Vibration* effects. After running our aggregation method, 10 time intervals containing *Wind* and 16 containing *Vibration* remained. Tables 9 and 10 show the effects of *Wind* and *Vibration* obtained from the crowd contributions compared to the sensory effect annotations with an intensity greater than or equal to 50% found in our reference dataset.

Much like in the case of the *Babylon A.D.* and *Earth* videos, workers did not notice events which in the reference dataset were associated with low-intensity tactile stimulus (*Wind* and *Vibration*) in *Formula 1* video. However, most of the time they noticed spans with an intensity stronger than 50%, which may indicate that less pronounced effects are more related to the expression of authorship than something highly expected by most viewers. At the beginning of the clip, drivers were slow and there was an indication of *Wind* in the public dataset; however, the workers were not of the same opinion, since they would feel the *Wind* in the environment of the movie. Furthermore, the workers spontaneously indicated *Vibration* when the cars accelerated, which was however not present in the reference dataset.

**Table 8** Earth - Wind

| Author | | Crowd | |
|---|---|---|---|
| start | end | start | end |
| 6.5 | 10.0 | 9.97 | 11.67 |
| 12.7 | 14.0 | 12.59 | 12.99 |
| 17.7 | 21.0 | 18.00 | 22.90 |
| 23.0 | 29.2 | 25.00 | 27.90 |
| 33.0 | 33.9 | 32.90 | 33.20 |
| 35.6 | 39.2 | 34.78 | 35.78 |
| | | 38.40 | 39.10 |
| 41.2 | 44.8 | | |
| | | 47.35 | 47.45 |
| | | 55.10 | 55.90 |
| | | 59.04 | 60.94 |
| | | 63.00 | 65.00 |

Besides, the crowd covered almost all *Vibration* effects annotated in this dataset even with low intensities.

The analysis of the results for this video shows that, predominantly, the crowd complemented the effects annotated by the author with *Vibration* effects for those events

**Table 9** Formula 1 - Vibration

| Author | | Crowd | |
|---|---|---|---|
| start | end | start | end |
| 7.67 | 7.97 | | |
| | | 11.60 | 13.40 |
| | | 16.00 | 18.90 |
| | | 23.60 | 28.20 |
| 31.0 | 32.3 | 31.51 | 31.91 |
| | | 33.59 | 36.29 |
| | | 36.68 | 37.58 |
| | | 43.82 | 44.02 |
| 47.0 | 47.9 | | |
| | | 54.03 | 56.83 |
| | | 62.00 | 66.00 |
| | | 71.16 | 71.76 |
| | | 78.75 | 78.95 |
| | | 88.95 | 91.65 |
| | | 92.43 | 96.53 |
| 96.7 | 99.0 | 98.21 | 101.21 |
| 100.7 | 101.2 | | |
| | | 102.01 | 102.71 |

**Table 10** Formula 1 - Wind

| Author | | Crowd | |
|---|---|---|---|
| start | end | start | end |
| 9.0 | 17.0 | 16.00 | 18.90 |
| 26.6 | 29.2 | 28.10 | 29.20 |
| 31.0 | 32.3 | | |
| | | 34.00 | 34.20 |
| 41.7 | 43.4 | | |
| | | 45.40 | 46.80 |
| 47.0 | 47.9 | | |
| | | 52.00 | 59.50 |
| | | 63.00 | 63.10 |
| 65.5 | 66.5 | | |
| | | 69.00 | 69.20 |
| 70.0 | 72.0 | | |
| 75.5 | 79.0 | | |
| 80.5 | 91.0 | 82.30 | 83.30 |
| 96.7 | 105.0 | 100.43 | 100.93 |
| 108.0 | 116.0 | 108.78 | 112.38 |

containing cars accelerating, and with *Wind* effects for those containing car overtaking scenes.

### 6.4 Crowdsourcing and mulsemedia content authoring

The experimental results show that our crowdsourcing approach for sensory effects authoring is a viable alternative when combined with an expert fine-tuning of mulsemedia authoring. While watching the annotated videos, we realized that most of the differences between the MPEG-V SEM from the dataset and that from the crowd could be justified. We believe that individuals are subject to natural bias and oversight when authoring mulsemedia content due to their unique prior experiences and the expected QoE of users is subject to what they are feeling.

A collective common sense to indicate time intervals of sensory effects can thus be an effective starting point for mulsemedia content annotation. On the other hand, it is still necessary to incorporate expert advice to fine-tune sensory effects attributes such as intensity, location, and so forth. This approach could be opportune for the industry to turn their multimedia videos into mulsemedia ones, outsourcing the hard work of pointing out the sensory effects presented in their library, and then, fine-tuning the work with their own experts.

**Table 11** Contributions collected and aggregated

| Task | Items | Contributions | Annotations | Filtered | Aggregated |
|---|---|---|---|---|---|
| 1 | 3 videos | 61 | 434 | 306 | 38 segments |
| 2 | 38 segments | 150 | 312 | 278 | 60 effects |
| Total | **41 items** | **211** | **746** | **584** | **98 items** |

**Table 12** Cost of the campaign

| Cost (USD) | |
|---|---|
| Per Contribution | 0.05 |
| Contributions | 10.55 |
| Platform Fees | 1.98 |
| Total | 12.53 |

Table 11 summarizes the number of contributions collected in each task, as well as the number of contributions remaining after the application of the filter criteria and the number of results produced by the aggregation method in each task.

In total, 211 contributions were collected during the entire process. Of these contributions, 746 annotations were obtained. Applying the reliability filters during the process, 584 annotations were considered valid, that is, there was a 78.28% effectiveness rate in the contributions.

The 211 contributions were each paid with 0.05 USD, resulting in an overall amount of 10.55 USD. Including crowdsourcing platform costs, the total spent on the campaign was 12.53 USD. These values are summarized in Table 12.

## 7 Conclusion

This paper introduces a completely new approach for authoring mulsemedia content based on crowdsourcing contributions. We compared our results to a public mulsemedia dataset to assess the proximity of the information provided by the crowd. The results pointed that the authoring made by the crowd adds to the public dataset and vice-versa.

A major observation of our study is that the effects identified by the crowd are largely not the same as those annotated by the author. This was already expected, since the effects obtained by the crowd reflect the workers' point of view regarding the intervals in which they believe that it makes sense to have sensory effects, whereas the author has a greater concern in transmitting his artistic point of view through of effects. The crowd was able to indicate the semantic associations related to the effects of *Wind* and *Vibration*, however, it was clear that the proposed insertions were for evident effects. The practical use of this is to adopt a hybrid approach in which the author divides the authoring with the multitude, delegating to the workers the work of identifying the insertions of greater intensity that relate to the user experience, allowing the author to concentrate on authoring more refined representation of his artistic vision. In other words, the authoring of the crowd did not replace that of the author's, but rather the help becomes more appropriate to improve the quality of the user experience.

For instance, in *Formula 1* the workers almost always realized *Wind* effects and *Vibration* when cars were accelerating. These effects were not authored in the reference dataset. This draws attention to the possibility of using the proposed approach to complement annotations made by experts.

Equally important, the idea behind our approach relies on the combination of the intuitive judgment of several individuals, common sense, and the refinement of an expert in order to take the best of each perspective to provide an alternative method for authoring mulsemedia content. As defined by van Holthoorn and Olson [36] *"common sense consists of knowledge, judgment, and taste which is more or less universal and which is held more or less*

*without reflection or argument."* The presented approach takes advantage of common sense emerged from the crowd in terms of expected sensory effects associated with each video scene. Also, it could be timely for mass production of coherent mulsemedia content without taking endless hours of an expert to start the process from scratch.

It is worth noticing that mulsemedia annotation does not automatically lead to mulsemedia authoring. For instance, in the case of *Wind* effects, where there is a lingering effect, the fact that the crowd didn't identify the *Babylon A.D.*'s segment [73.8, 74.6] s does not necessarily mean that the fan has to be switched on/off at these points but it is a cue. Indeed, because of lingering effects, network and device delays, a propagation delay should be considered.

Another important observation is that even with a limited number of contributions, the crowd nonetheless associated sensory effects for the most evident situations, such as when explosions, gunshots or car accelerations occur in scenes. Hence, this approach could be applied to build larger datasets storing video annotated with sensory effects. Moreover, these datasets could be used for training systems based on machine learning to detect the previous situations in AV contents automatically. In addition, our approach can also be used for QoE evaluation purposes in a manner akin to that of Yue, Wang and Cheng [41].

Future work includes finding how to use the wisdom of the crowd to collect fine-tuned attributes to the maximum extent as well as the automatic generation of MPEG-V metadata for mulsemedia content without the need for incorporating traditional annotation tools in the process. Furthermore, in a similar fashion to this work, the annotation of other types of effects (e.g. olfactory and thermal) using the cascade crowdsourcing process could well be included in future experiments.

**Acknowledgements** This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES), the Brazilian National Council for Scientific and Technological Development (CNPq), and the Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES). Estêvão Bissoli Saleme thankfully acknowledges support from the Federal Institute of Espírito Santo. Prof. Gheorghita Ghinea gratefully acknowledges funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 688503 for the NEWTON project (http://www.newtonproject.eu).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

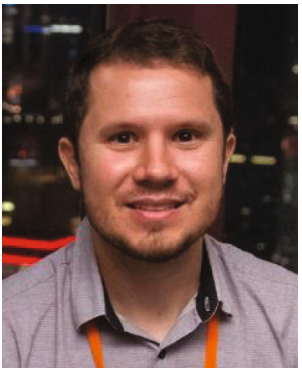
## References

1. Ademoye OA, Murray N, Muntean GM, Ghinea G (2016) Audio masking effect on inter-component skews in olfaction-enhanced multimedia presentations. ACM Trans Multimedia Comput Commun Appl 12(4):51:1–51:14. https://doi.org/10.1145/2957753
2. Amorim MN, Neto FRA, Santos CAS (2018) Achieving complex media annotation through collective wisdom and effort from the crowd. In: 2018 25th international conference on systems, signals and image processing (IWSSIP). IEEE, pp 1–5. https://doi.org/10.1109/IWSSIP.2018.8439402
3. Ballan L, Bertini M, Del Bimbo A, Seidenari L, Serra G (2011) Event detection and recognition for semantic annotation of video. Multimedia Tools Appl 51(1):279–302. https://doi.org/10.1007/s11042-010-0643-7
4. Bartocci S, Betti S, Marcone G, Tabacchiera M, Zanuccoli F, Chiari A (2015) A novel multimedia-multisensorial 4d platform. In: AEIT International annual conference (AEIT), 2015. IEEE, pp 1–6. https://doi.org/10.1109/AEIT.2015.7415215
5. Chen J, Yao T, Chao H (2018) See and chat: automatically generating viewer-level comments on images. MTAP: Multimedia Tools Appl, 1–14. https://doi.org/10.1007/s11042-018-5746-6

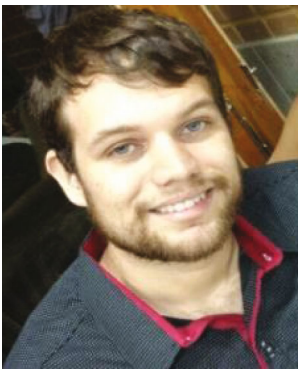
6. Cho H (2010) Event-based control of 4d effects using mpeg rose. Master's thesis, School of Mechanical, Aerospace and Systems Engineering, Division of Mechanical Engineering. Korea Advanced Institute of Science and Technology. Master's Thesis
7. Choi B, Lee ES, Yoon K (2011) Streaming media with sensory effect. In: 2011 international conference on information science and applications (ICISA). IEEE, pp 1–6. https://doi.org/10.1109/ICISA.2011.5772390
8. Chowdhury SN, Tandon N, Weikum G (2016) Know2look: commonsense knowledge for visual search. In: Proceedings of the 5th workshop on automated knowledge base construction, pp 57–62
9. Covaci A, Zou L, Tal I, Muntean GM, Ghinea G (2018) Is multimedia multisensorial?-a review of mulsemedia systems. ACM Comput Survey (CSUR) 51(5):91
10. Cross A, Bayyapunedi M, Ravindran D, Cutrell E, Thies W (2014) Vidwiki: enabling the crowd to improve the legibility of online educational videos. In: Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing. ACM, pp 1167–1175
11. Di Salvo R, Spampinato C, Giordano D (2016) Generating reliable video annotations by exploiting the crowd. In: IEEE Winter conf. on applications of computer vision (WACV). IEEE, pp 1–8. https://doi.org/10.1109/WACV.2016.7477718
12. Dumitrache A, Aroyo L, Welty C, Sips RJ, Levas A (2013) A.: "dr. detective": combining gamification techniques and crowdsourcing to create a gold standard in medical text. 16–31
13. Egan D, Brennan S, Barrett J, Qiao Y, Timmerer C, Murray N (2016) An evaluation of heart rate and electrodermal activity as an objective qoe evaluation method for immersive virtual reality environments. In: 8th international conference on quality of multimedia experience (qoMEX'16). https://doi.org/10.1109/QoMEX.2016.7498964
14. Foncubierta Rodríguez A, Müller H (2012) Ground truth generation in medical imaging: a crowdsourcing-based iterative approach. In: Proceedings of the ACM multimedia 2012 workshop on crowdsourcing for multimedia, CrowdMM '12. ACM, New York, pp 9-14. https://doi.org/10.1145/2390803.2390808
15. Galton F (1907) Vox populi (the wisdom of crowds). Nature 75(7):450–451
16. Ghinea G, Timmerer C, Lin W, Gulliver SR (2014) Mulsemedia: State of the art, perspectives, and challenges. ACM Trans Multimedia Comput Commun Appl 11(1s):17:1–17:23. https://doi.org/10.1145/2617994
17. Gottlieb L, Choi J, Kelm P, Sikora T, Friedland G (2012) Pushing the limits of mechanical turk: qualifying the crowd for video geo-location. In: Proceedings of the ACM multimedia 2012 workshop on crowdsourcing for multimedia. ACM, pp 23–28
18. Hardman L, Obrenović Ž, Nack F, Kerhervé B, Piersol K (2008) Canonical processes of semantically annotated media production. Multimedia Syst 14(6):327–340. https://doi.org/10.1007/s00530-008-0134-0
19. Kim S, Han J (2014) Text of white paper on mpeg-v. Tech. Rep ISO/IEC JTC 1/SC 29/WG 11 W14187, San Jose, USA
20. Kim SK (2013) Authoring multisensorial content. Signal Process Image Commun 28(2):162–167. https://doi.org/10.1016/j.image.2012.10.011
21. Kim SK, Yang SJ, Ahn CH, Joo YS (2014) Sensorial information extraction and mapping to generate temperature sensory effects. ETRI J 36(2):224–231. https://doi.org/10.4218/etrij.14.2113.0065
22. Lasecki W, Miller C, Sadilek A, Abumoussa A, Borrello D, Kushalnagar R, Bigham J (2012) Real-time captioning by groups of non-experts. In: Proceedings of the 25th annual ACM symposium on User interface software and technology - UIST '12, UIST '12. ACM Press, New York, pp 23-33. https://doi.org/10.1145/2380116.2380122
23. Masiar A, Simko J (2015) Short video metadata acquisition game. In: 10th international workshop on semantic and social media adaptation and personalization (SMAP). IEEE, pp 61–65. https://doi.org/10.1109/SMAP.2015.7370092
24. McNaney R, Othman M, Richardson D, Dunphy P, Amaral T, Miller N, Stringer H, Olivier P, Vines J (2016) Speeching: mobile crowdsourced speech assessment to support self-monitoring and management for people with parkinson's. In: Proceedings of the 2016 CHI conference on human factors in computing sys - CHI '16, CHI '16. ACM Press, New York, pp 4464-4476. https://doi.org/10.1145/2858036.2858321
25. Murray N, Lee B, Qiao Y, Muntean GM (2016) The influence of human factors on olfaction based mulsemedia quality of experience. https://doi.org/10.1109/QoMEX.2016.7498975
26. Neto FRA, Santos CAS (2018) Understanding crowdsourcing projects: a systematic review of tendencies, workflow, and quality management. Inf Process Manag 54(4):490–506. https://doi.org/10.1016/j.ipm.2018.03.006

27. Oh HW, Huh JD (2017) Auto generation system of mpeg-v motion sensory effects based on media scene. In: 2017 IEEE international conference on consumer electronics (ICCE). IEEE, pp 160–163. https://doi.org/10.1109/ICCE.2017.7889269
28. Rainer B, Waltl M, Cheng E, Shujau M, Timmerer C, Davis S, Burnett I, Ritz C, Hellwagner H (2012) Investigating the impact of sensory effects on the quality of experience and emotional response in web videos. In: 4th international workshop on quality of multimedia experience (qoMEX). IEEE, pp 278–283. https://doi.org/10.1109/QoMEX.2012.6263842
29. Sadallah M, Aubert O, Prié Y (2014) Chm: an annotation- and component-based hypervideo model for the web. Multimed Tools Appl 70(2):869–903. https://doi.org/10.1007/s11042-012-1177-y
30. Saleme EB, Celestrini JR, Santos CAS (2017) Time evaluation for the integration of a gestural interactive application with a distributed mulsemedia platform. In: Proceedings of the 8th ACM on multimedia systems conference, MMSys'17. ACM, New York, pp 308-314. https://doi.org/10.1145/3083187.3084013
31. Saleme EB, Santos CAS, Ghinea G (2018) Coping with the challenges of delivering multiple sensorial media. IEEE MultiMedia, 1–1. https://doi.org/10.1109/MMUL.2018.2873565
32. Shin SH, Ha KS, Yun HO, Nam YS (2016) Realistic media authoring tool based on mpeg-v international standard. In: 2016 8th international conference on ubiquitous and future networks (ICUFN). IEEE, pp 730–732. https://doi.org/10.1109/ICUFN.2016.7537133
33. Taborsky E, Allen K, Blanton A, Jain AK, Klare BF (2015) Annotating unconstrained face imagery: a scalable approach. In: International conference on biometrics (ICB). IEEE, pp 264–271. https://doi.org/10.1109/ICB.2015.7139094
34. Teki S, Kumar S, Griffiths TD (2016) Large-scale analysis of auditory segregation behavior crowdsourced via a smartphone app. PLos ONE, 11(4). https://doi.org/10.1371/journal.pone.015
35. Timmerer C, Waltl M, Rainer B, Hellwagner H (2012) Assessing the quality of sensory experience for multimedia presentations. Signal Process Image Commun 27(8):909–916. https://doi.org/10.1016/j.image.2012.01.016
36. van Holthoon F, Olson D (1987) Common sense: the foundations for social science. Common sense. University Press of America, Lanham
37. Waltl M, Rainer B, Timmerer C, Hellwagner H (2013) An end-to-end tool chain for sensory experience based on mpeg-v. Signal Process Image Commun 28(2):136–150. https://doi.org/10.1016/j.image.2012.10.009
38. Waltl M, Timmerer C, Hellwagner H (2010) Improving the quality of multimedia experience through sensory effects. In: Second international workshop on quality of multimedia experience (qoMEX). IEEE, pp 124–129
39. Waltl M, Timmerer C, Rainer B, Hellwagner H (2012) Sensory effect dataset and test setups. In: 4th international workshop on quality of multimedia experience (qoMEX). IEEE, pp 115–120. https://doi.org/10.1109/QoMEX.2012.6263841
40. Yuan Z, Bi T, Muntean GM, Ghinea G (2015) Perceived synchronization of mulsemedia services. IEEE Trans Multimedia 17(7):957–966. https://doi.org/10.1109/TMM.2015.2431915
41. Yue T, Wang H, Cheng S (2018) Learning from users: a data-driven method of qoe evaluation for internet video. MTAP: Multimedia Tools Appl, 1–32. https://doi.org/10.1007/s11042-018-5918-4
42. Zhai H, Lingren T, Deleger L, Li Q, Kaiser M, Stoutenborough L, Solti I (2013) Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processing. J Med Internet Res 15(4):1–17. https://doi.org/10.2196/jmir.2426

**Marcello Novaes de Amorim** is a doctorate candidate in the Computer Science Department at Federal University of Espírito Santo, Brazil. He received the graduation degree in Computer Science from UFES, Brazil, in 2005, and the M.Sc. degree in Computer Science from the Federal University of Espírito Santo, Brazil, in 2007. His research interests include multimedia systems, human computation and crowdsourcing. Contact him at novaes@inf.ufes.br.



**Estêvão Bissoli Saleme** is currently Ph.D. candidate in the Computer Science at Federal University of Espírito Santo (UFES), Brazil. From August 2018 and Februay 2019, he was an Academic Visitor at Brunel University London, UK. He received the B.Sc. degree in Information Systems from FAESA, Brazil, in 2008, and the M.Sc. degree in Computer Science from the UFES, in 2015. His current research interests include multimedia/mulsemedia systems, middlewares and frameworks, interactive multimedia, media transport and delivery. Contact him at estevaobissoli@gmail.com.



**Fábio Ribeiro de Assis Neto** earned the B.Sc. and M.Sc. degrees in Computer Science from the Federal University of Espírito Santo (UFES), Brazil, in 2012 and 2017, respectively. His research interests include crowdsourcing, multimedia systems, and human computation. Contact him at fabio.ribeiro.neto@gmail.com.

**Dr. Celso A. S. Santos** is a Professor in the Department of Informatics at Federal University of Espírito Santo (UFES), Brazil. He received the B.S. degree in Electrical Engineering from UFES in 1991, and the M.S. degree in Electrical Engineering (Electronic Systems) from the Polytechnic School of the University of São Paulo, Brazil, in 1994. In 1999, he received his Dr. degree at Informatique Fondamentalle et Parallelisme from Universitè Paul Sabatier de Toulouse III, France. His recent research interests focus on multimedia/mulsemedia systems and applications, synchronization, and crowdsourcing systems. Contact him at saibel@inf.ufes.br.

**Dr. Gheorghita Ghinea** is a Professor in the Computer Science Department at Brunel University, United Kingdom. He received the B.Sc. and B.Sc. (Hons) degrees in Computer Sci-ence and Mathematics, in 1993 and 1994, respectively, and the M.Sc. degree in Computer Science, in 1996, from the University of the Witwatersrand, Johannesburg, South Africa; he then received the Ph.D. degree in Computer Science from the University of Reading, United Kingdom, in 2000. His work focuses on building adaptable cross-layer end-to-end communication systems incorporating user multisensorial and perceptual requirements. He is a member of the IEEE and the British Computer Society. Contact him at george.ghinea@brunel.ac.uk.

## Affiliations

**Marcello Novaes de Amorim[1] · Estêvão Bissoli Saleme[1] · Fábio Ribeiro de Assis Neto[1] · Celso A. S. Santos[1] · Gheorghita Ghinea[2]**

Estêvão Bissoli Saleme
estevaobissoli@gmail.com

Fábio Ribeiro de Assis Neto
fabio.ribeiro.neto@gmail.com

Celso A. S. Santos
saibel@inf.ufes.br

Gheorghita Ghinea
george.ghinea@brunel.ac.uk

1 Federal University of Espírito Santo, Av. Fernando Ferrari, 514, 29075-910 - Vitória-ES, Brazil

2 Brunel University London, Wilfred Brown Building 215, Kingston Lane, Middlesex UB8 3PH, Uxbridge, England