

QoE of Cross-modally Mapped Mulsemmedia: an Assessment Using Eye Gaze and Heart Rate

GEBREMARIAM MESFIN, *Brunel University London, United Kingdom*

NADIA HUSSAIN, *Brunel University London, United Kingdom*

ELAHE KANI-ZABIHI, *University of West London, United Kingdom*

ALEXANDRA COVACI, *University of Kent, United Kingdom*

ESTÊVÃO B. SALEME, *Federal University of Esp rito Santo, Brazil*

GHEORGHITA GHINEA¹, *Brunel University London, United Kingdom*

Abstract. A great deal of research effort has been put in exploring crossmodal correspondences in the field of cognitive science which refer to the systematic associations frequently made between different sensory modalities (e.g. high pitch is matched with angular shapes). However, the possibilities cross-modality opens in the digital world have been relatively unexplored. Therefore, we consider that studying the plasticity and the effects of crossmodal correspondences in a mulsemmedia setup can bring novel insights about improving the human-computer dialogue and experience. Mulsemmedia refers to the combination of three or more senses to create immersive experiences. In our experiments, users were shown six video clips associated with certain visual features based on color, brightness, and shape. We examined if the pairing with crossmodal matching sound and the corresponding auto-generated haptic effect, and smell would lead to an enhanced user QoE. For this, we used an eye-tracking device as well as a heart rate monitor wristband to capture users' eye gaze and heart rate whilst they were experiencing mulsemmedia. After each video clip, we asked the users to complete an on-screen questionnaire with a set of questions related to smell, sound and haptic effects targeting their enjoyment and perception of the experiment. Accordingly, the eye gaze and heart rate results showed significant influence of the cross-modally mapped multisensorial effects on the users' QoE. Our results highlight that when the olfactory content is crossmodally congruent with the visual content, the visual attention of the users seems shifted towards the correspondent visual feature. Crossmodally matched media is also shown to result in an enhanced QoE compared to a video only condition.

Keywords: Mulsemmedia, crossmodal correspondence, quality of experience, gaze tracking, heart rate

1 INTRODUCTION

Multisensory experiences have mostly been studied in the psychology field where interactions between smell and taste have been explored. Multisensory integration occurs between two or more sensory modalities including touch, sound, vision, smell and taste. Mulsemmedia (Multiple Sensorial Media) incorporates more than visual and audio information, it includes new media types such as haptics, olfaction and gustatory. It has led to new opportunities as well as challenges in research, academia, industry, and for immersive technologies [18, 67].

Authors' addresses: Gebremariam Mesfin, Brunel University London, United Kingdom, gebremariam.assres@brunel.ac.uk; Nadia Hussain, Brunel University London, United Kingdom, nadia.hussain@brunel.ac.uk; Elahe Kani-Zabihi, University of West London, United Kingdom, Elahe.Kani@uwl.ac.uk; Alexandra Covaci, University of Kent, United Kingdom, a.covaci@kent.ac.uk; Est v o B. Saleme, Federal University of Esp rito Santo, Brazil, estevaobissoli@gmail.com; Gheorghita Ghinea, Brunel University London, United Kingdom, george.ghinea@brunel.ac.uk.

¹ Corresponding Author

In this emerging field, there have been several explorations on the practicality and possibility of integrating different media types into applications. Thanks to the advent of novel technologies and innovative devices to artificially produce sensory effects along with systems able to deliver this kind of experience to the users [56] the addition of multiple sensory effects has been essential to improving immersion and *presence* in the user's environment.

Ghinea et al. [17] believe it can be achieved by the user's sensation perception, classifying it as a result of a complex set of processes in which biological sensors send structured electrical signals to the brain (except for specific chemoreceptors), which in turn, frame unconscious sensations patterns. Thereby, they help to determinate whether an upcoming sensory input is authentic. Additionally, Möller and Raake [40] consider that perception goes through two stages before being completely realized: (i) conversion of stimuli through the respective sensory organ into neural signals, and (ii) processing and transmission of these neural signals from the central nervous system to the cerebral cortex, resulting in specific perceptions in the person's perceptual world. It all will permeate what is called Quality of Experience (QoE) of users.

QoE stems from the combination of the achievement of users' expectations regarding the utility, the level of enjoyment considering their personalities, and their current state [4]. Users exposed to multisensory experiences have reported a noticeable increase in QoE [27, 41, 46, 48, 75, 78, 79]. Although there have been studies in the cognitive and digital world with regard to the perception of individual senses, there are hitherto unsettled questions when it comes to crossmodal correspondences. In crossmodal correspondence, a stimulus in one modality can be associated with another. For example, in the non-digital world, the smell of lemon and high pitch audio can be associated to sharp objects [21, 63]. However, it is not yet clear whether or not the multisensorial effect of the component modalities generated out of such crossmodal associations would enhance the users' QoE in the digital world.

In this article, we report on an experiment designed to explore whether cross-modally mapped multisensorial effects (olfaction, sound, and auto-generated haptic) from visual features of videos enhance the users' QoE. We hypothesize that taking into account crossmodal mappings whilst creating mulsemmedia systems could lead to more immersive and effective experiences for the users.

This article is organized as follows. Section 2 brings related work, focusing on auditory-visual crossmodal correspondences research in psychology, computer graphics and human computer interaction, mulsemmedia and QoE. Section 3 presents the user study on QoE in crossmodal mulsemmedia. Section 4 depicts the results and discusses the work. Finally, Section 5 provides a concluding summary and underlines topics for future investigation.

2 RELATED WORK

2.1 Auditory-visual and olfactory-visual crossmodal correspondences research in psychology

Past experiences shape unconscious sensations patterns, which in turn, will influence the way humans feel the upcoming experiences. Thus, a new stimulus in one modality might be associated with another one; for instance, the pitch in audition can be associated to visual features like brightness. Outside the digital world, crossmodal correspondences have been observed between different sensory modalities such as visual, sound, touch, smell, and taste [7, 15, 57, 59, 62, 76].

Non-arbitrary crossmodal correspondences mappings between auditory and visual stimuli have been found through experimental approaches in simple stimulus dimensions such as loudness and brightness, as well as in more complex stimuli such as shapes/images and words. Marks [37] detected an association between lighter colors with higher pitches and louder sounds. Besides, sound has also

been linked to other compound characteristics such as shapes. In the same study, Marks [37] gathers an evidence that high pitched tones are related to angular shapes and low-pitch sounds are connected to rounder shapes. Hagtvedt and Brasel [20] found an association between the frequency of music and the lightness of a colored object. With the help of an eye-tracker, they concluded that visual attention was steered in the direction of light-colored objects under the influence of high-frequency sounds. It makes evident that sound can be employed to exploit users' attention.

Nonetheless, over the last decades, researchers have started to document the existence of crossmodal correspondences also between olfactory and visual stimuli. For instance, in [19], the authors provided one of the first examples of olfactory-visual correspondences, showing that there are strong correlations between odors and colors. Bergamot smell was associated with yellow, cinnamon with red, pine with green, etc. In [31], the authors investigated how color lightness varies with perceived odor intensity and found an inverse correlation. Pleasantness and quality of odors were also analyzed in studies, such as [53, 65]. In [11], the authors investigated the robustness of these crossmodal associations for a random sequence of odors (strawberry vs. spearmint) and color patches (pink vs. turquoise) and found these correspondences both systematic and robust. In [57] authors took a different approach and investigated the crossmodal associations between the abstract symbols designed for the representation of an odor and the correspondent odor. They showed that the matching exists and is mediated by hedonic valence of cues. In [10], participants were asked to select a color they were associating with an odor. They observed that when odors were described in abstract terms it was less likely to find a color match, while when the participants were describing the odor with a source-based term ("smells like banana") their color choices reflected more accurately the odor source. This and other studies like [26, 61] show that the mechanisms underlying these associations could be related to semantics, emotions or natural co-occurrence.

If synaesthesia is unidirectional, crossmodal correspondences are bidirectional: e.g., hearing high-pitched sound is matched with small objects and seeing small objects is paired with high-pitched sounds. The fact that crossmodal correspondences are bidirectional might mean that at least some of the crossmodal correspondences are also transitive, which is again different from synaesthesia [12]. Though the multidimensionality of the precepts at stake seems to indicate the possibility to predict the relationship between different attributes, transitivity should not be expected in every case. For instance, we know that louder sounds correspond to bigger objects and that lower pitch corresponds to larger size, thus louder sounds should correspond to higher pitch. However, this was not observed in related studies [12].

2.2 Auditory-visual and olfactory-visual crossmodal correspondences research in computer graphics and human computer interaction

There has been little work related to crossmodal correspondences between visual and auditory media beyond the area of cognitive sciences. The studies of Mastoropoulou et al. [39] and Mastoropoulou [38] on the effect of auditory stimuli over visual perception pointed out that when only sound emitting objects are delivered in high quality and the rest of the scene in lower quality, the visual quality is not impacted.

In [3], the authors focused on different senses for investigating crossmodal correspondences: sight and olfaction. They found out that the scent of fresh cut grass can distract viewers from the task of identifying the animation quality (flyover of a grass terrain). Hulusić et al. [24] aimed at discovering the influence of beat rates in static scenes. They found out that lower beat rates impact the perception of low frame rates. Thereafter, they investigated how camera movement speed and the sounds influence the smoothness of the animation [25]. Ramic-Brkic et al. [49] were concerned about how viewers perceive the graphics' quality in the presence of distinct modalities such as auditory, olfactory, and ambient

temperature. What they realized was that strong perfume, high temperature, and audio noise have an influence on the users' perceived rendering quality. Apart from selective rendering, Tanaka and Parkinson [35] studied the crossmodal mapping between digital audio and the haptic domain dedicated to audio producers with visual impairments. To do so, they created a device called Haptic Wave, an input/output interface that renders audio data as kinesthetic information. In [35], the authors explored the impact of audio on haptic to improve the quality of eating for denture users. They built a device to increase the food texture using sound. Ranasinghe et al. [51] applied crossmodal perception to create Vocktail, a system to introduce flavor as a digitally controllable media involving color, smell, and taste modalities. In [23], the authors found associations between sweetness and red rounded shapes, and sourness and green angular shapes with a fast animation speed in the literature. Then, they also found out that specific combinations of visualizations and animation types have an influence on yogurt's taste perception. Tag et al. [69] explored cross-modal correspondence between haptic and audio output for meditation support. The goal of the haptic/audio design was to guide the user into a particular rhythm of breathing. In [28], the authors discuss the effect of scented material on physical creations showing that odor-shape correspondence exists in an active, free association creation session. Moreover, it also indicates the potential of using crossmodal correspondences for HCI in the design of future interactive experiences.

The multisensory user experience is also a semiotic process [29] and designing for it can take different stances depending on the experimental goals. Positive emotional outcome is dependent on the context of the design and its appraisal is strongly connected to multisensory integration. Expectations have an important role in HCI, thus crossmodal correspondences could be one of the underlying dynamics of a positive experience [52]. As can be seen, studies on crossmodal correspondences research in computer graphics and human-computer interaction provide insights about sensory replacement/composition under different circumstances. These mappings have a promising potential in designing interfaces and displays that tap into a user's mental model [71]. Thus, we believe that crossmodal mappings could reveal insightful information in other contexts to help to understand the users' perception and therefore improve human-computer interaction.

2.3 Mulsemmedia and QoE

There has been an increasing interest in creating multimedia applications augmented with media on top of the traditional audio-video (AV) content [18]. They aim at stimulating other senses beyond sight and hearing such as touch [14], smell [16] or taste [50, 51] with the aim to increase the user's QoE and to explore novel methods for interaction [43]. Therefore, the term mulsemmedia refers to the use of at least three different media types, that is, multimedia and at least one non-traditional media [18].

Mulsemmedia systems generally undergo a workflow for (i) production, (ii) distribution, and (iii) rendering [6]. First, different sensory effects metadata are produced or automatically generated in synchronization with an AV content. This process can be performed by a human or acquired through various sensors (e.g. camera, microphone, motion capture) that capture real-world information, or synthesized using computers (e.g. a virtual 3D space in a game) [55]. Many tools have been developed to aid this process, such as SEVino [74], SMURF [32], RoSE Studio [5], and Real 4D studio [58]. The works of Kim et al. [33] and Oh and Huh [47] are endeavors to automatically produce mulsemmedia metadata. Although haptic effects can be captured [9], making a reliable and lasting record of taste and smell from the real world is still a challenge.

Following that, the mulsemmedia effects can be encoded for transport, processed and emitted for distribution to providers, distributed to the end-users and then decoded by systems, and finally, rendered by different devices, which in turn, will deliver them to the end users. Mulsemmedia players and renderers

to be used with other multimedia applications have also been created to reproduce and deliver mulsemmedia experiences, notably SEMP [74] and PlaySEM [54], which are open-source. A mulsemmedia system entails weaving multiple technologies to connect different entities, distribute the sensory signals, and render sensory effects appropriately Saleme et al. [55]. Whilst developing mulsemmedia systems, it is crucial to have ways to deliver different sensory content consistently as well as of paramount importance to be aware of the challenges that might arise when delivering mulsemmedia [56]. The main motivation behind adding mulsemmedia components is to augment the level of immersion and QoE of users [43].

QoE is defined as the level of delight or displeasure a user feels whilst experiencing an application or a service in computers taking into account mainly subjective measures such as their personalities and current state. It can be assessed either by conducting subjective surveys [2, 75, 77, 78] or objective evaluation [13, 30]. In addition, technical recommendations have been used together such as ITU-R-BT.500-13, ITU-TBT.500, and ITU-T-P.910. Therefore, mulsemmedia systems' evaluations can lead to a high degree of qualitative differentiation in terms of QoE. Although objective evaluations are low-cost and carried out faster than subjective ones, they might put researchers on the wrong track if they consider just a few parameters. For instance, researchers should know if the user has some heart-related problems before they measure the user's heart rate because it can lead to misleading conclusions. Thus, taking current emotional states into consideration from different perspectives could reveal useful insights. The work of Egan et al. [13] is a sample of the combination of objective and subjective QoE evaluations. They correlated the results of both and found out that high values for heart rate and electrodermal activity had to do with physiological arousal- one of the factors associated with user QoE. Another work [30], showed the potential and benefits of using these objective metrics as indicators of user QoE for immersive experiences in augmented reality applications. Indeed, if used appropriately, physiological measures can be useful in affective state monitoring, chiefly in a multimodal setup [34].

By satisfying users' expectations and incrementing the levels of utility/enjoyment of applications or services, mulsemmedia has not only contributed directly to QoE, but also indirectly such as presented in the studies of Yuan et al. [78], Yuan et al. [79], and Ademoye et al. [2]. They have pointed out that mulsemmedia can partially mask an AV sequence's decreased quality as well as synchronization skews, thus enhancing the user's perceived QoE. Furthermore, mulsemmedia has the capacity to aid memory [1], to improve virtual realism, to more easily convey information between physical and digital environments [80], and to contribute to pattern recognition [66].

The question of how to improve the user experience in immersive systems is still an open one. Adding sensory modalities seems to be a reasonable way according to the literature. However, it is also relevant to pay attention to crossmodal correspondences, which have seldom been considered when designing mulsemmedia systems although our perceptual experiences are affected by them. Very little is known about the combination of senses in the digital world and what occurs as soon as one stimulus is stronger than the others. Indeed, crossmodal interactions could be handy when it comes to getting over a specific sensory deprivation or situational impairment such as to see or feel something in darkness [22]. Given this, mulsemmedia appears as a prospective scenario to develop the knowledge on crossmodal correspondences hitherto limited to setups on traditional multimedia. By understanding crossmodality applied to mulsemmedia systems, this comprehension could be also beneficial to prepare effective mulsemmedia experiences.

3 USER STUDY: QUALITY OF EXPERIENCE IN CROSSMODAL MULSEMEDIA

The experiments we designed are aimed to investigate the potential influence of using crossmodal correspondences concepts in designing mulsemmedia on the QoE experienced by the users. More specifically, we used six videos characterized by dominant visual features: color (blue, yellow), brightness (low, high), shape (round, angular). Participants viewed these videos enhanced with crossmodally matching sound while wearing a haptic vest with vibration motors. We chose to use the vibrotactile display because literature has shown that participants exhibit an increased emotional response to media with haptic enhancement [72].

3.1 Participants

Twelve participants (7 males, 5 females) took part in the experiment and were randomly assigned to either one of an equal-sized Experimental (EG) or Control Group (CG), respectively. Users were aged between 18-41 years old and hailed from diverse nationalities and educational backgrounds (undergraduate and postgraduate students as well as academic staff). All participants spoke English and self-reported as being computer literate.

3.2 Experimental apparatus

The videos were displayed on a computer monitor with a resolution of 1366x768 pixels, and a viewing area of 1000x700 pixels in the center of the screen. An EyeTribe eye tracker controlled by a custom written Java code was employed to record eye-gaze patterns on a Windows 10 Laptop with 8GB RAM powered by an Intel Core i5 processor. The viewing screen was placed between 45-75 cm from the eyes of the participants, as this was the recommended distance for EyeTribe calibration². We chose to use the EyeTribe eye tracker because this was demonstrated to be accurate enough in studies on gaze points and fixations [8].

Participants sat in a chair without armrests facing the screen. All participants wore i-shine³ headphones, a vibrotactile KOR-FX⁴ gaming vest, and a Mio Link heart rate wristband⁵. To facilitate the vibrotactile experience we chose the KOR-FX gaming vest that utilizes 4DFX based acousto-haptic signals to enable haptic feedback to the upper chest and shoulder regions. The vest is wirelessly connected to a control box meant to accept the standard sound output of the sound card of a computer.

The olfactory emitting device was provided by the Exhalia SBi4⁶, which was considered by previous research more reliable and more robust than existing devices [44]. This was placed at 0.5m from the assessor, allowing her/him to detect the smell in 2.7-3.2 s, as shown in [43]. The SBi4 can store up to four interchangeable scent cartridges at a time, but we used a single slot in our experiments to prevent the mixing of scents. These cartridges contain scented polymer through which air is blown (through four built-in-fans). The synchronized presentation of the olfactory data was controlled through a program built using Exhalia's Java-based SDK. Users of this type of devices obtain additional information about environmental factors while becoming more immersed/involved in their experience [42]. A snapshot of the experimental setup is shown in Fig. 1.

² <http://theyetribe.com/dev.theyetribe.com/dev.theyetribe.com/start/index.html>

³ <https://www.ishine-trade.com/Headphones-Earphones>

⁴ <http://korfx.com/products>

⁵ <https://www.mioglobal.com/>

⁶ <https://www.exhalia.com/>

3.3 Audio Visual Olfactory content

As illustrated in Table 1 there were six videos selected based on their dominant visual features such as color, brightness and angularity of objects. The olfactory content consisted of six scents: bergamot, lily, clear lavender (low intensity), lavender (high intensity), lemon and raspberry. All videos in our experiment were 120 seconds long. For the EG, the audio was adjusted to a frequency of 328Hz (high pitch condition) and 41Hz (low pitch condition).

The accompanying auditory and olfactory content was modified in line with principles of auditory-visual and olfactory-visual crossmodal correspondences that were previously shown in the literature. The video with dominant yellow images (V1) was watched accompanied by high pitch sounds and bergamot odor, while the one dominantly blue (V2) by low pitch sounds and lily odor [19, 60, 68].



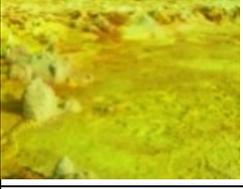
Fig. 1. Experimental setup. The users were wearing: (1) i-shine headphone, (2) the KOR-FX haptic vest, their eye gaze was captured with (3) the EyeTribe eye tracker, while their heart rate was measured with (4) Mio Link; olfactory effects were diffused using Exhalia (5).

In V3, where brightness was considered the dominant visual cue, low pitch sounds and low intensity lavender odor were delivered concurrently to the users, while in V4, where the brightness was high, the auditory content consisted of high pitch sounds and the olfactory content of high intensity lavender odor, based on [19, 36]. Finally, V5, the video displaying angular shapes, was matched with high pitch sounds and lemon odor, whilst V6, where the dominant shape was round, was delivered with low pitch sounds and raspberry odor [21, 63].

3.4 Procedure

Pre-experiment study. Before the experiments, we carried out a small pilot study with two participants to get feedback on their thoughts and experience while trying our system. This was aimed to give us feedback on the experimental process and research instruments employed. Since participants reported that the high pitch audio volume was loud, we lowered its intensity to enhance user comfort during the experiment.

Table 1. Snapshots from the six videos used during the experiment with their themes, dominant visual cues and the conditions for the EG in each case. The CG experienced only visual content, without any type of crossmodally generated content (olfactory, auditory or vibrotactile).

Video Snapshot	Description
<p style="text-align: center;">V1</p> 	<p>Theme. Waves Visual cue. Color: Blue EG. Auditory: Low pitch, Olfactory: Lillial CG. Only visual content</p>
<p style="text-align: center;">V2</p> 	<p>Theme. Sulphur springs Visual cue. Color: Yellow EG. Auditory: High pitch, Olfactory: Bergamot CG. Only visual content</p>
<p style="text-align: center;">V3</p> 	<p>Theme. Solar eclipse Visual cue. Brightness: Low EG. Auditory: Low pitch, Olfactory: Lavender (low intensity) CG. Only visual content</p>
<p style="text-align: center;">V4</p> 	<p>Theme. Sunrise upon the Arctic Visual cue. Brightness: High EG. Auditory: High pitch, Olfactory: Lavender (high intensity) CG. Only visual content</p>
<p style="text-align: center;">V5</p> 	<p>Theme. Fly over San Francisco Visual cue. Shape: Angular EG. Auditory: High pitch, Olfactory: Lemon CG. Only visual content</p>
<p style="text-align: center;">V6</p> 	<p>Theme. Bouncing balls Visual cue. Shape: Round EG. Auditory: Low pitch, Olfactory: Raspberry CG. Only visual content</p>

Conditions. There were two conditions that differed in the provided content:

- (1) In the **experimental condition** (associated with the EG) users were exposed to altered audio (modified pitch) which it matched the corresponding dominant visual features. The dominant visual cue was also accompanied by crossmodally correspondent olfactory cues.
- (2) In the **placebo condition** (carried out by the CG), the users were only exposed to the visual content. Thus, although they wore headphones and a haptic vest and the fan of the olfactory device was running, no type of content (auditory, vibrotactile nor olfactory) was distributed to users.

Eye-tracking calibration. At the beginning of the experiment, participants underwent an eye-tracking calibration exercise in which they were asked to focus on 9 equally spaced points situated on a 3x3 grid. Participants were randomly divided in two groups of 6 each and watched the six videos in a random order for both EG and CG. All participants used the devices identified in Fig. 1. The experimental sessions were conducted individually and lasted between 24 to 37 minutes.

Collected data. For each participant we collected two objective measures:

- Gaze points - as a measure of visual attention and interest. These were collected as a set of (x,y) pixel co-ordinates, with a sampling frequency of 30 Hz, matching the frame rate of the videos.
- Heart rate - as a measure of user emotional arousal whilst experiencing the system. The Mio Link wristband consists of an optical heart rate module (OHRM) that utilizes photoplethysmography (PPG) to measure continuous heart rate alongside an accelerometer unit to measure and correct for movement artifacts [73]. Accelerometer data assessing a user's movement is entered into an algorithm that compensates for movement artifacts in the optical signal. The raw data provided comprised heart rate readings sampled once every second.

Table 2. Self-reported QoE questions.

Item	Description
Q1	The smell was relevant to the video clip I was watching.
Q2	The smell came across strong.
Q3	The smell was distracting.
Q4	The smell was consistent with the video clip when released.
Q5	The smell was annoying.
Q6	The smell faded away slowly after watching the video clip.
Q7	The smell enhanced my viewing experience.
Q8	The sound was related to the video clip I was watching.
Q9	The sound came across loud.
Q10	The sound was distracting.
Q11	The sound was annoying.
Q12	The sound enhanced the sense of reality whilst watching the video clip.
Q13	The sound enhanced my viewing experience.
Q14	I enjoyed watching the video clip whilst wearing a Haptic Vest.
Q15	The Haptic Vest effects were relevant to the video clip I was watching.
Q16	The vibration was distracting.
Q17	The vibration was annoying.
Q18	The Haptic Vest effects enhanced the sense of reality whilst watching the video clip.
Q19	The Haptic Vest effects enhanced my viewing experience.
Q20	Overall, I enjoyed the multisensorial experience.

Participants also completed a subjective questionnaire (Table 2) at the end of the experiment. Each

question was answered on a 5-item Likert scale, anchored at one end with “Strongly Disagree” and with “Strongly Agree” at the other.

4 RESULTS AND DISCUSSION

In this section, we present analysis and discussion of results of the data obtained from eye-tracker, heart-rate monitor, and on-screen QoE questionnaire (Table 2). Data were analyzed with the IBM Statistical Package for the Social Sciences (SPSS) for Windows version (release 23.0). An ANalysis Of VAriance (ANOVA), suitable to test the significant differences of three or more categories, as well as one sample t-test and independent sample t-test, suitable to check whether a sample mean is statistically different from a hypothesized population mean, and, respectively, to identify significant differences between two categories [64], were applied to analyze the participants’ responses. A significance level of $p < 0.05$ was adopted for the study.

4.1 Analysis of eye-gaze data

The eye gaze data was collected at a sampling rate of about the same as the frame rate and hence we obtained a total of 3600 eye gaze (30 eye gaze/sec x 120 sec) locations per each video clip. As mentioned in Section 3.2, the viewing area for the videos measures 1000x700 pixels and it is centered on a 1366x768 pixels screen.

For analysis purposes, this viewing area is partitioned in 20 equal segments across both the X and Y axes, resulting in a total of 400 eye gaze cells of 50x35 pixels each. For each such cell of a particular video frame, we first counted the number of individuals, in the CG and EG respectively, whose eye gaze fell into it. We then calculated, for each video frame, the summation of the absolute differences in eye gaze count between the EG and CG across all cells, as shown in equation (1).

$$\sum_{i=1}^N |\Delta Gaze_i|, \text{ where } 1 < i < N, N = 400 \text{ viewing cells/frame} \quad (1)$$

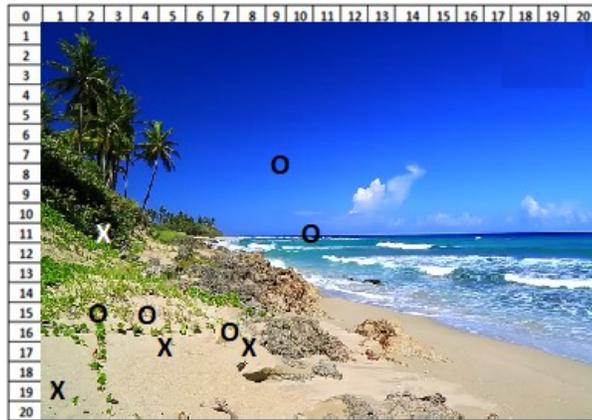


Fig. 2. Points where the participants gazed at the 50th frame of video V1 (X ∈ EG, O ∈ CG).

In this regard, the minimum and maximum eye gaze difference count between the EG and CG are $\text{Min } \Delta = 0$ and $\text{Max } \Delta = 12$, respectively. For example, Figure 2 shows the eye gaze count at the 50th frame of video 1 observed from participants in both CG and EG.

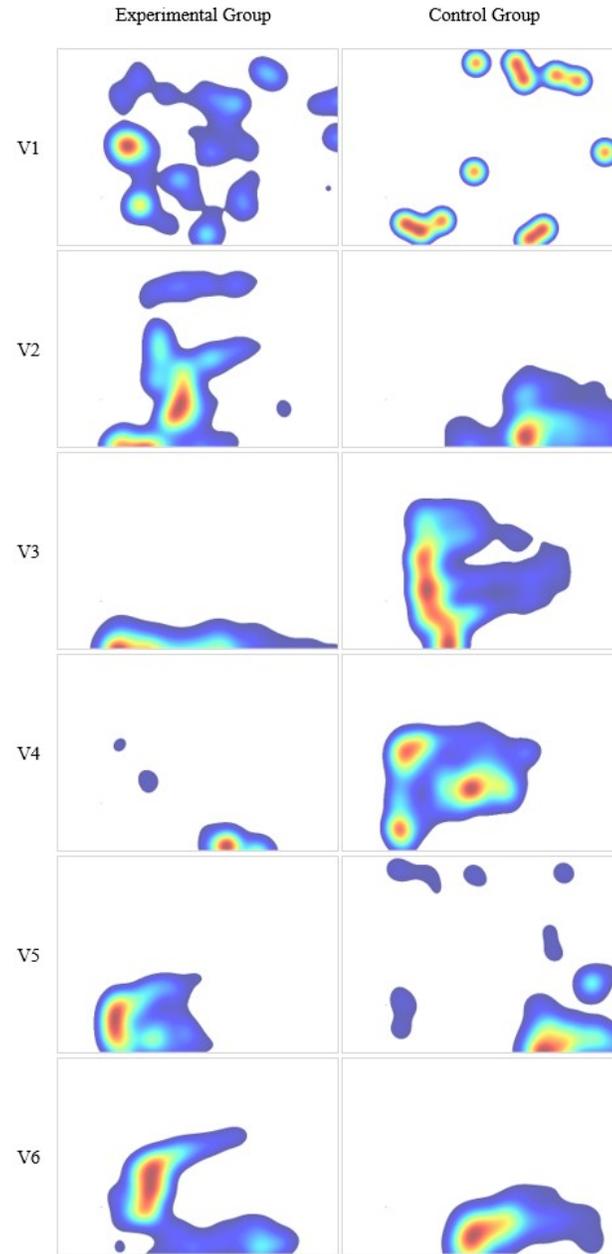


Fig. 3. General heat map across the video clips. Red means most viewed and most fixated on. Yellow refers to some views, but less fixation. Green indicates less views and fixations. Blue suggests least viewed and hardly any fixations. White indicates hardly any views and no fixations.

The eye gaze data for all the videos is represented in heat maps in Figure 3. This is split into EG (on the left side) and CG (on the right side). The videos are sequenced in rows from V1 to V6. As can be seen in V1, the EG seemed to explore the scenario whereas the CG focused in diversified points. In contrast, EG participants had broader scan patterns in V2. In V3 and V4, which contain low brightness and high brightness respectively, the EG focused on the lower part of the viewing area where white standouts, although most of the times V3 presents a dark area. V5 presents the angular shapes in dynamic sequences, which means they were spread out. Here, CG participants examined the video with more dispersed gaze patterns compared to the EG. The heat map suggested that the latter was more focused when exposed to angular shapes, high pitch, and lemon. Finally, in V6, both groups focused their attention on the circular shapes in different positions on the screen.

In order to analyze the eye gaze data, a one sample t-test of the eye gaze difference count was performed and is shown in Table 3. The result reveals that there are statistically significant differences in eye gaze between the EG and CG for all the six videos ($p < 0.05$). However, as the difference between the groups was the audio soundtrack (the CG had no soundtrack, whilst the EG had a mapped high/low pitch sound), and smell effects (the CG had no smell whilst the EG had congruent smell), we cannot deduce that the difference in eye gaze count is due to the difference in between groups in the experienced audio (and haptic effect), smell or both. Thus, further analysis is provided in the subsequent sections to identify the impact of each.

Table 3. One sample t-test of eye gaze difference count.

Video	Sig. (2-tailed)	Mean difference	95% Confidence	
			Lower	Upper
V1	.000	-3.724	-3.75	-3.70
V2	.000	-5.005	-5.03	-4.98
V3	.000	-5.284	-5.31	-5.26
V4	.000	-4.622	-4.65	-4.60
V5	.000	-4.754	-4.78	-4.73
V6	.000	-5.700	-5.72	-5.68

4.2 Analysis of heart ratedata

As a physiological metric, we employed heart rate data which was collected at the rate of one reading per second and measured in beats per minute (bpm). Accordingly, we collected 120 heart rate readings for each video. The heart rate readings from the CG varied between 60bpm and 100bpm whilst, for the EG, these ranged between 75bpm and 110bpm, with the means for each video illustrated in Figure 4.

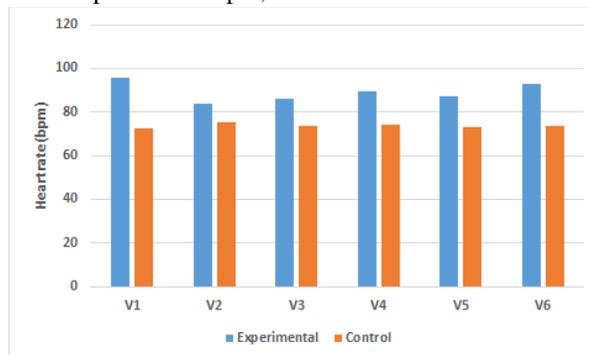


Fig. 4. Average heart rate data for all video clips

In Figure 5 we present the mean heart rate gathered every second for each of the six videos in both CG and EG. We observe a tendency for a higher heart rate in the EG for the whole duration of the videos. In order to understand whether this tendency is statistically significant, we undertook an independent samples t-test, the results of which are shown in Table 4. The results in Table 4 evidence a statistically significant difference between the heart rates of the two groups for all the videos. This indicates that the two groups experienced a different mood in the two setups: (i) the one using crossmodally matching sound and smell (EG) and (ii) the one where no sound and smell were provided to the visual dominant features (CG). We remind the reader that the sound served as an input also for the vibrotactile feedback.

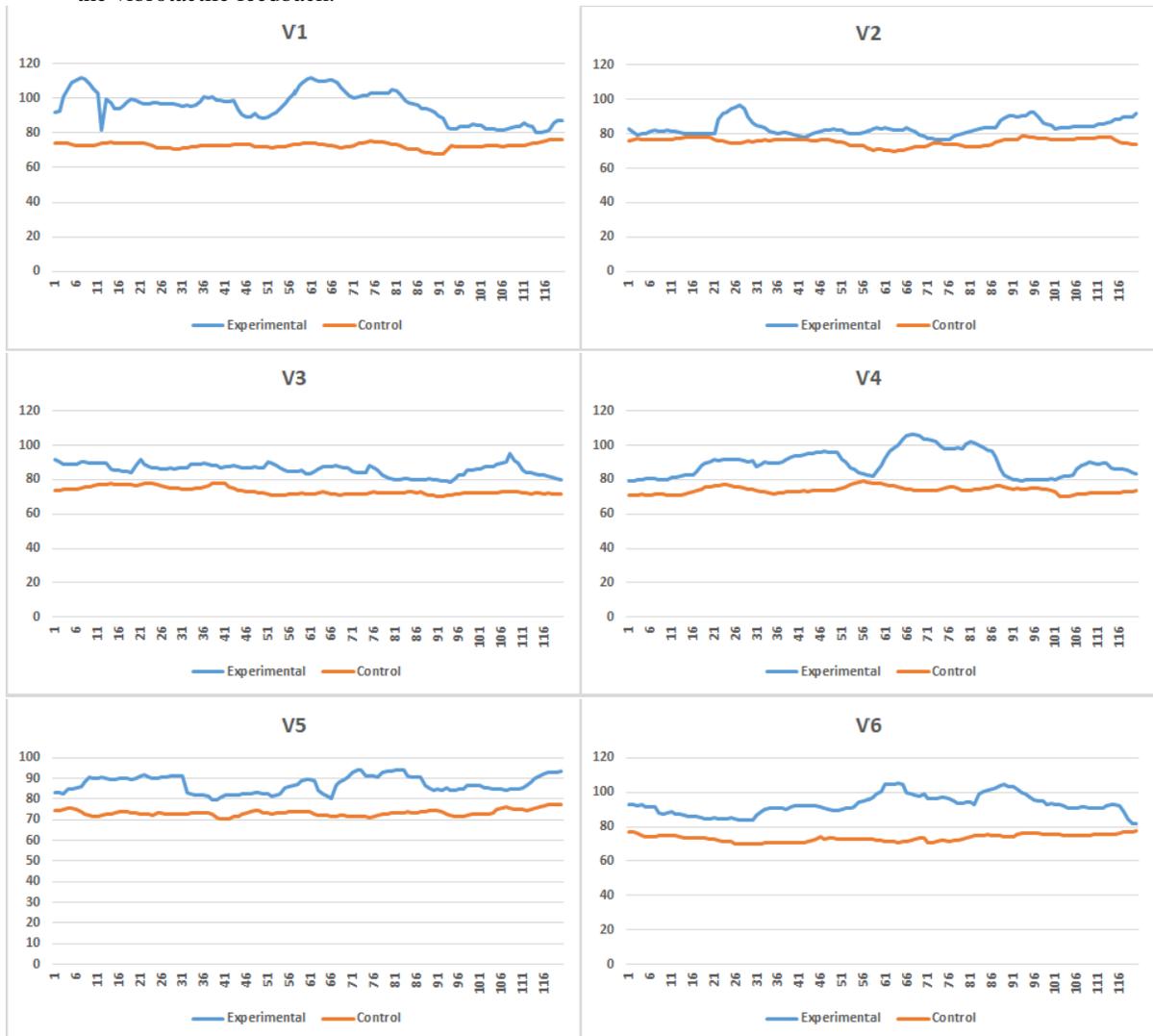


Fig. 5. Average heart rate data (bpm) of the participants for each video.

4.3 Analysis of self-reported QoE

Participants self-reported QoE by answering a series of 20 Likert scale questions, as shown in Table 2. For analysis, we converted the scores of each negatively-phrased questions (Q2, Q3, Q5, Q6, Q9, Q10, Q11, Q16, and Q17) to the equivalent score associated with a positively-phrased counterpart.

Initially, we performed a two way ANOVA with group type and video type as independent variables and the responses to the 20 QoE questions as the dependent variables, the results of which are presented in Table 5. As can be seen, there is a statistically significant difference between the EG and CG (Group) for all questions except Q3, Q5, Q14, and Q17; and the difference in QoE between the videos (Video) is statistically insignificant. Table 5 also shows that the interaction of the independent variables (Group*Video) has generally statistically insignificant effect for all questions on the self-reported QoE (dependent variable) except for Q15. Accordingly, a Post Hoc Tukey test analysis was conducted on all questions (except Q15), which also resulted in statistically insignificant values.

Table 4. Independent samples t-test of heart rate data.

Video	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval	
				Lower	Upper
V1	.000	23.061	.841	21.404	24.718
V2	.000	08.260	.456	07.360	09.159
V3	.000	12.518	.368	11.793	13.243
V4	.000	15.700	.718	14.286	17.114
V5	.000	13.996	.389	13.230	14.761
V6	.000	19.371	.553	18.281	20.461

Table 5. ANOVA multivariate test result for each question.

Source	Dep. Var.	df	F	Sig.	Source	Dep. Var.	df	F	Sig.	Source	Dep. Var.	df	F	Sig.
Group	Q1	1	158.401	.000	Video	Q1	5	.259	.933	Group*Video	Q1	5	.259	.933
	Q2	1	16.425	.000		Q2	5	.279	.923		Q2	5	.123	.987
	Q3	1	.255	.616		Q3	5	.483	.788		Q3	5	.206	.959
	Q4	1	12.712	.001		Q4	5	.395	.850		Q4	5	.780	.568
	Q5	1	.048	.828		Q5	5	.679	.641		Q5	5	.134	.984
	Q6	1	12.859	.001		Q6	5	.321	.898		Q6	5	1.813	.124
	Q7	1	11.463	.001		Q7	5	.979	.438		Q7	5	1.080	.381
	Q8	1	34.845	.000		Q8	5	1.299	.276		Q8	5	1.536	.192
	Q9	1	23.937	.000		Q9	5	.507	.770		Q9	5	.480	.790
	Q10	1	15.123	.000		Q10	5	.437	.821		Q10	5	1.049	.397
	Q11	1	8.485	.005		Q11	5	.831	.533		Q11	5	1.602	.173
	Q12	1	15.961	.000		Q12	5	1.113	.363		Q12	5	.916	.477
	Q13	1	15.805	.000		Q13	5	.293	.915		Q13	5	.702	.624
	Q14	1	.890	.349		Q14	5	.235	.946		Q14	5	.655	.659
	Q15	1	18.211	.000		Q15	5	1.060	.391		Q15	5	2.398	.048
	Q16	1	11.598	.001		Q16	5	.113	.989		Q16	5	.124	.987
	Q17	1	3.882	.053		Q17	5	.346	.883		Q17	5	.553	.736
	Q18	1	18.728	.000		Q18	5	.798	.556		Q18	5	.832	.532
	Q19	1	11.523	.001		Q19	5	.415	.837		Q19	5	.933	.466
	Q20	1	5.738	.020		Q20	5	.384	.858		Q20	5	.193	.964

The mean and standard deviation in the self-reported QoE is 3.07 and 1.18 for the EG, respectively; and 2.91 and 1.16 for the CG, respectively. Farther explanation corresponding to each of the questions with respect to the results in Table 5 and Figure 6 is presented next.

- In the case of **Q1**, the mean response is significantly higher (2.69) in EG than the CG. This implies that respondents have noticed the relevance of the various smells rendered for the respective video clips.
- In **Q2**, the average is significant (3.83) in the CG which shows that there was intensity variation in the rendering of the smell effect across the video clips.
- The mean of **Q3** is slightly higher (statistically insignificant) in the CG which means that the smells were generally less distracting.
- In **Q4**, the mean response is significantly higher (3.22) in the EG than the CG. This implies that the smell was consistent across the videos.
- The average value of **Q5** for EG is slightly higher (3.25) than the CG. This implies that the smells were perceived quite pleasant.
- In **Q6**, the mean is significantly higher in the CG which means that the lingering effect of the smells was noticeable as compared to the CG.
- The mean response corresponding to **Q7** is significantly higher (3.25) in EG than the CG which means that the smell effects (congruent smells) have significant contribution to the overall QoE when viewing the video clips.
- In **Q8**, the mean response is significant (3.31) in the EG than the CG. This indicates that respondents have noticed the relevance of the high/low pitch audios for the respective video clips.
- The mean response for **Q9** significantly higher (4.03) in the CG which means that there was noticeable loudness variations of the sound across the video clips.
- In **Q10**, the mean is significant (3.78) in the CG which shows that the high/low pitched sound were generally less distracting.
- The mean of **Q11** for the CG is significantly higher (3.64) than the EG. This implies that the high/low pitched sound were generally found to be not annoying by experimental participants.
- In **Q12**, the average response is significantly higher (3.11) in the EG than the CG. This means that the high/low pitched sound (which were congruent to the visual features of the video clips) has triggered a sense of reality that significantly enhances the overall QoE.
- The mean answer for **Q13** is significant (3.17) in the EG which signifies that the sound effect contributed to the overall QoE when viewing the video clips.
- In **Q14**, the average response of EG is slightly higher (3.31) than the CG. This denotes that the haptic effects which were automatically generated out of the content-congruent sound have contributed to the enjoyment.
- The mean score of the EG in **Q15** is significant (2.86) which shows that respondents have noticed the relevance of the haptic effect for the respective videos.
- In **Q16**, the mean QoE is significantly higher (3.69) in the CG than the EG which means that the vibrations on the chest while wearing the haptic vest had certain distracting effects.
- The mean of **Q17** is slightly higher (3.61) in the CG. This implies that the haptic effects generated out of the high/low pitched sound were generally not significantly annoying.

- In **Q18**, the mean is significant (3.08) in the EG which indicates that the haptic effect (generated out of the high/low pitched sound which is congruent to the visual features of the videos) has significantly enhanced the sense of reality while watching the video clips.
- The mean value corresponding to **Q19** is significantly higher (3.25) in EG than the CG. This means that the haptic effects generated out of the content-congruent sound have significant contribution to the overall QoE when viewing the video clips.
- In **Q20**, the mean is significantly higher (3.86) in the EG than the CG. This implies that the combined multisensorial effect of the content-congruent smell, sound, and the auto-generated haptic has contributed to the enjoyment while watching the video clips.

Because the interaction of the independent variables (Group*Video) for **Q15** showed a statistically significant value, we conducted simple main effect analysis (Table 6). Thus, V3 and V5 showed statistically significant lower scores obtained from the EG compared to their CG counterparts ($F(1,60) = 12.140, p < .05$ and $F(1,60) = 14.448$, respectively) which implies that the haptic effects generated out of the content-congruent sound was significantly less relevant to the video clips having more dark and angular features than the other four video clips. However, in the case of V1, V2, V4, and V6, differences in participant scores between the two groups were not significant.

Table 6. Simple main effects analysis (Q15).

Video	df	F	Sig.	Mean Difference (Experimental-Control)
V1	1	.903	.346	-0.500
V2	1	.100	.753	-0.167
V3	1	12.140	.001	-1.833
V4	1	2.508	.119	-0.833
V5	1	14.448	.000	-2.000
V6	1	.100	.753	-0.167

The results corresponding to most of the self-reported QoE questions indicated that the content-congruent smell, sound, and the auto-generated haptic effects have enhanced the users' QoE while watching the video clips. This is substantiated by the mean responses of the EG and CG for all the questions (3.07 and 2.91, respectively); and the statistically significant difference values corresponding to most of the questions in (Table 5) which implies that the cross-modally mapped (overall) multisensorial setting has enhanced the QoE.

In general, our analysis of the difference in eye gaze count (Table 3) and heat map of the eye gaze patterns (Figure 3) showed that the cross-modally mapped multisensorial effects have significantly influenced the users' perception. Significantly high heart rate recording is also observed due to the introduction of multisensorial effects in the EG of participants (Table 4, Figure 5). Additionally, analysis of the self-reported QoE evidenced the eye gaze and heart results revealing that the multisensorial effects involving content-congruent high/low pitch sound, smell, and haptic have significantly enhanced the QoE.

The findings also indicate that the positive impact of multisensorial effects on users' QoE is substantiated by integrating cross-modally mapped component effects in a mulsemmedia context. This implies there exists a noticeable cross-modal correspondence in a digital world between the visual features of videos and audio pitches which substantiates studies in [19, 60, 68]. Similarly, such correspondence exists between the visual features of the videos and smell effects [19, 21, 60, 63, 68].

5 CONCLUSIONS

This paper presents an exploratory study that begins to establish how crossmodal correspondences could be systematically explored for multisensory content design. In our study, we examined the impact of crossmodal mappings between visual features and auditory media, and visual features and olfactory media on user QoE. These mappings were previously shown to be favorable to design interfaces and displays that tap into users' mental model leading to more immersive and effective experiences [81].

By employing multimedia video clips, eye tracker, haptic vest and heart rate monitor wristband in our experiment, we gathered results from both subjective surveys and objective metrics. The use of the eye tracker exposed that there were significant differences in both EG and CG. Gaze heat maps showed that the EG was more focused when experiencing mulsemmedia, except when exposed to the combination of yellow, high pitch and bergamot smell. Although we cannot draw strong conclusions based on the gaze patterns of the participants, we observe that when the olfactory content is crossmodally congruent with the visual content, the visual attention of the users seems shifted towards the correspondent visual feature (e.g., exploration and focus on the blue sky for V1; wider exploration area for the round shapes (more balls) for V6).

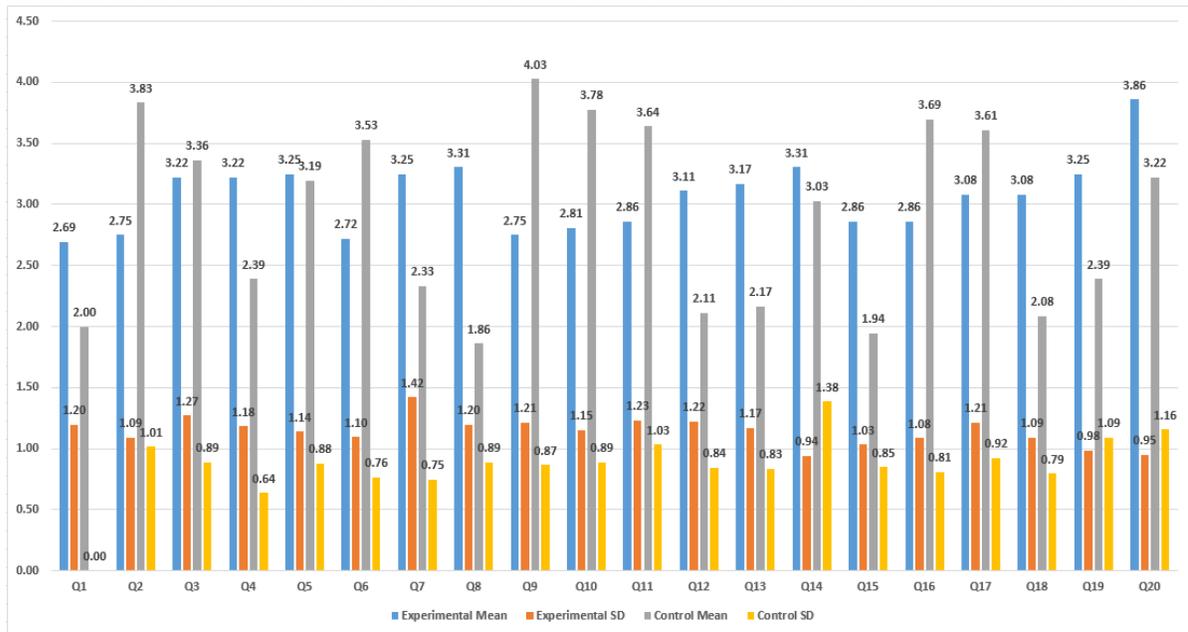


Fig. 6. Average QoE for the EG andCG.

The heart rate responses were also significant. This could be due to users experiencing different moods, not only that the heart rate was much higher in the EG as opposed to the CG. One of the possible reasons could be that the use of high vs low pitch may have affected the users' viewing experience, whereas in the CG there was no sound limiting the immersion as well as the experience. By reflecting on both groups, it shows that the use of sound and smell did have a positive effect and increased users QoE to a certain degree.

The self-reported responses support the eye gaze and heart rate results, revealing that the multisensory effects involving crossmodally mapped (content-congruent) smell, sound (high/low

pitch), and auto-generated haptic have enhanced the QoE compared to a visual only condition. This also implies that there exists a noticeable cross-modal correspondence from visual features to audio pitches and smell effects.

Overall, our results might be indicative of causality between visual attention and the presence of additional content that matches the dimensions meant to be attended, but further work needs to be done in order to validate this. Indeed, one of the limitations of this study is that it does not look into differences between the effects of content created using crossmodal principles and other types of multisensory content (e.g., where correspondences are semantic). Thus, although we show that the attention and the QoE benefit from the multisensory content, it is not obvious if this is caused by employing crossmodal principles. Another limitation of the study is the relatively small number of participants, which makes it unclear how our findings would generalise in other setups. Also, the study reported here is an exploratory one, which has raised many interesting paths for future investigation. Among these, worthy of mention are repeating the experiment when users view videos with other, non-coherent (neutral) stimuli as well as when viewing content with non-congruent stimuli. All are valuable future pursuits. Further work could also be done to explore what content is more appealing to users. Categorizing the content into different topics and carrying out a pilot study amongst few users will provide us with what types of media content they would prefer to watch. Moreover, odors influence mood, work performance, and many other forms of behavior and this has been evidenced in our study. We intend to further investigate in the future by comparing original sound with altered high and low pitch as well as looking at employing additional, different odors for crossmodal matching.

ACKNOWLEDGMENTS

This paper was funded by the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 688503. Estêvão B. Saleme thankfully acknowledges support from IFES and CAPES (process no. 88881.187844/2018-01).

REFERENCES

1. O. A. Ademoye and G. Ghinea. 2013. Information recall task impact in olfaction-enhanced multimedia. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 9, 3 (2013), 17.
2. O. A. Ademoye, N. Murray, G-M. Muntean, and G. Ghinea. 2016. Audio masking effect on inter-component skews in olfaction-enhanced multimedia presentations. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 12, 4 (2016), 51.
3. B. R. Brkic, A. Chalmers, K. Boulanger, S. Pattanaik, and J. Covington. 2009. Cross-modal affects of smell on the real-time rendering of grass. In *Proceedings of the 25th Spring Conference on Computer Graphics*. ACM, 161–166.
4. K. Brunnström, S. A. Beker, K. De Moor, A. Doms, S. Egger, M-N. Garcia, T. Hossfeld, S. Jumisko-Pyykkö, C. Keimel, M-C. Larabi, et al. 2013. Qualinet white paper on definitions of quality of experience. (2013).
5. B. Choi, E. S. Lee, and K. Yoon. 2011. Streaming media with sensory effect. In *Information Science and Applications (ICISA), 2011 International Conference on*. IEEE, 1–6. <https://doi.org/10.1109/ICISA.2011.5772390>
6. Covaci, L. Zhou, I. Tal, G-M. Muntean, and G. Ghinea. 2018. Is Multimedia Multisensorial? - A Review of Mulsemmedia Systems. *ACM Computing Surveys (CSUR)* (2018), 35 pages. In press. <https://bura.brunel.ac.uk/bitstream/2438/16449/1/Fulltext.pdf>
7. A-S. Crisinel and C. Spence. 2009. Implicit association between basic tastes and pitch. *Neuroscience letters* 464, 1 (2009), 39–42.
8. Dalmaijer. 2014. Is the low-cost EyeTribe eye tracker any good for research? *PeerJ PrePrints* (2014).
9. Danieau, A. Lécuyer, P. Guillotel, J. Fleureau, N. Mollet, and M. Christie. 2013. Enhancing audiovisual experience with haptic feedback: a survey on HAV. *IEEE Transactions on Haptics* 6, 2 (2013), 193–205.
10. J. M. de Valk, E. Wnuk, J. L. A. Huisman, and A. Majid. 2017. Odor-color associations differ with verbal descriptors for odors: A comparison of three linguistically diverse groups. *Psychonomic Bulletin & Review* 24, 4 (01 Aug 2017), 1171–1179. <https://doi.org/10.3758/s13423-016-1179-2>
11. M. L. Demattè, D. Sanabria, and C. Spence. 2006. Cross-Modal Associations Between Odors and Colors. *Chemical Senses* 31, 6 (2006), 531–538. <https://doi.org/10.1093/chemse/bjj057>

12. O. Deroy and C. Spence. 2013. Why we are not all synesthetes (not even weakly so). *Psychonomic Bulletin & Review* 20, 4 (01 Aug 2013), 643–664. <https://doi.org/10.3758/s13423-013-0387-2>
13. D. Egan, S. Brennan, J. Barrett, Y. Qiao, C. Timmerer, and N. Murray. 2016. An evaluation of Heart Rate and ElectroDermal Activity as an objective QoE evaluation method for immersive virtual reality environments. In *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*. IEEE, 1–6.
14. M. Eid, J. Cha, and A. El Saddik. 2008. HugMe: A haptic videoconferencing system for interpersonal communication. In *Virtual Environments, Human-Computer Interfaces and Measurement Systems, 2008. VECIMS 2008. IEEE Conference on*. IEEE, 5–9.
15. D. Gal, S. C. Wheeler, and B. Shiv. 2007. Cross-modal influences on gustatory perception. (2007).
16. G. Ghinea and O. Ademoye. 2012. The sweet smell of success: Enhancing multimedia applications with olfaction. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 8, 1 (2012), 2.
17. G. Ghinea, F. Andres, and S. R. Gulliver. 2011. *Multiple Sensorial Media Advances and Applications: New Developments in MulSeMedia*. Information Science Reference.
18. G. Ghinea, C. Timmerer, W. Lin, and S. R. Gulliver. Mulsemmedia: State of the art, perspectives, and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.*, 11(1s):17:1–17:23, October 2014.
19. N. Gilbert, R. Martin, and S. E. Kemp. 1996. Cross-modal correspondence between vision and olfaction: the color of smells. *The American journal of psychology* (1996), 335–351.
20. H. Hagtvedt and S. A. Brasel. 2016. Cross-Modal Communication: Sound Frequency Influences Consumer Responses to Color Lightness. *Journal of Marketing Research* 53, 4 (2016), 551–562.
21. G. Hanson-Vaux, A-S. Crisinel, and C. Spence. 2012. Smelling shapes: Crossmodal correspondences between odors and shapes. *Chemical Senses* 38, 2 (2012), 161–166.
22. E. Hoggan and S. A. Brewster. 2006. Mobile crossmodal auditory and tactile displays. In *Proceedings of HAID 2006: First International Workshop on Haptic and Audio Interaction Design*. 9–12.
23. G. Huisman, M. Bruijnes, and D. K. Heylen. 2016. A moving feast: effects of color, shape and animation on taste associations and taste perceptions. In *Proceedings of the 13th International Conference on Advances in Computer Entertainment Technology*. ACM, 13.
24. V. Hulusic, G. Czanner, K. Debatista, E. Sikudova, P. Dubla, and A. Chalmers. 2009. Investigation of the beat rate effect on frame rate for animated content. In *Proceedings of the 25th Spring Conference on Computer Graphics*. ACM, 151–159.
25. V. Hulusic, K. Debatista, V. Aggarwal, and A. Chalmers. 2010. Exploiting audio-visual cross-modal interaction to reduce computational requirements in interactive environments. In *Games and Virtual Worlds for Serious Applications (VS-GAMES), 2010 Second International Conference on*. IEEE, 126–132.
26. M. Jacquot, F. Noel, C. Velasco, and C. Spence. 2016. On the Colours of Odours. *Chemosensory Perception* 9, 2 (01 Jun 2016), 79–93. <https://doi.org/10.1007/s12078-016-9209-z>
27. L. Jalal, M. Anedda, V. Popescu, and M. Murrioni. 2018. QoE Assessment for IoT-Based Multi Sensorial Media Broadcasting. *IEEE Transactions on Broadcasting* 64, 2 (June 2018), 552–560. <https://doi.org/10.1109/TBC.2018.2823914>
28. O. Jezler, E. Gatti, M. Gilardi, and M. Obrist. 2016. Scented Material: Changing Features of Physical Creations Based on Odors. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. ACM, New York, NY, USA, 1677–1683. <https://doi.org/10.1145/2851581.2892471>
29. J. P. P. Jokinen. 2015. Emotional User Experience and Feeling of Control. In *Proceedings of the 19th International Academic Mindtrek Conference (AcademicMindTrek '15)*. ACM, New York, NY, USA, 163–166. <https://doi.org/10.1145/2818187.2818284>
30. C. Keighrey, R. Flynn, S. Murray, and N. Murray. 2017. A QoE evaluation of immersive augmented and virtual reality speech & language assessment applications. In *Quality of Multimedia Experience (QoMEX), 2017 Ninth International Conference on*. IEEE, 1–6.
31. S. E. Kemp and A. N. Gilbert. 1997. Odor intensity and color lightness are correlated sensory dimensions. *The American journal of psychology* 110, 1 (1997), 35.
32. S-K. Kim. 2013. Authoring multisensorial content. *Signal Processing: Image Communication* 28, 2 (2013), 162–167. <https://doi.org/10.1016/j.image.2012.10.011>
33. S-K. Kim, S-J. Yang, C. H. Ahn, and Y. S. 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>
34. S. Koelstra, C. Muhl, M. Soleymani, J-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. 2012. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing* 3, 1 (2012), 18–31.
35. N. Koizumi, H. Tanaka, Y. Uema, and M. Inami. 2011. Chewing jockey: augmented food texture by using sound based on the cross-modal effect. In *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology*. ACM, 21.
36. L. E. Marks. 1974. On associations of light and sound: The mediation of brightness, pitch, and loudness. *The American journal of psychology* (1974), 173–188.

37. Lawrence E Marks. 1987. On cross-modal similarity: Auditory–visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance* 13, 3 (1987), 384.
38. G. Mastoropoulou. 2007. *The effect of audio on the visual perception of high-fidelity animated 3d computer graphics*. Ph.D. Dissertation. University of Bristol.
39. G. Mastoropoulou, K. Debattista, A. Chalmers, and T. Troscianko. 2005. Auditory bias of visual attention for perceptually-guided selective rendering of animations. In *Proceedings of the 3rd international conference on Computer graphics and interactive techniques in Australasia and South East Asia*. ACM, 363–369.
40. S. Möller and A. Raake (Eds.). 2014. *Quality of Experience: Advanced Concepts, Applications and Methods*. Springer, Cham. <https://doi.org/10.1007/978-3-319-02681-7>
41. J. Monks, A. Olaru, I. Tal, and G. M. Muntean. 2017. Quality of experience assessment of 3D video synchronised with multisensorial media components. In *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*. 1–6. <https://doi.org/10.1109/BMSB.2017.7986129>
42. G. Munster, T. Jakel, D. Clinton, and E. Murphy. 2015. Next mega tech theme is virtual reality. *gene* 612 (2015), 303–6452.
43. N. Murray, O. A. Ademoye, G. Ghinea, and G-M. Muntean. 2017a. A Tutorial for Olfaction-based Multisensorial Media Application Design and Evaluation. *ACM Computing Surveys (CSUR)* 50, 5 (2017), 67.
44. N. Murray, B. Lee, Y. Qiao, and G-M. Muntean. 2014. Multiple-Scent Enhanced Multimedia Synchronization. *ACM Trans. Multimedia Comput. Commun. Appl.* 11, 1s, Article 12 (Oct. 2014), 28 pages. <https://doi.org/10.1145/2637293>
45. N. Murray, B. Lee, Y. Qiao, and G. M. Muntean. 2016. The influence of human factors on olfaction based mulsemedia quality of experience. In *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. 1–6. <https://doi.org/10.1109/QoMEX.2016.7498975>
46. N. Murray, B. Lee, Y. Qiao, and G-M. Muntean. 2017b. The Impact of Scent Type on Olfaction-Enhanced Multimedia Quality of Experience. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 47, 9 (Sept 2017), 2503–2515. <https://doi.org/10.1109/TSMC.2016.2531654>
47. H-W. Oh and J-D. Huh. 2017. Auto generation system of MPEG-V motion sensory effects based on media scene. In *Consumer Electronics (ICCE), 2017 IEEE International Conference on*. IEEE, 160–163. <https://doi.org/10.1109/ICCE.2017.7889269>
48. B. Rainer, M. Walzl, E. Cheng, M. Shujau, C. Timmerer, S. Davis, I. Burnett, C. Ritz, and H. Hellwagner. 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'12)*. http://www-itec.uni-klu.ac.at/bib/files/QoMEX2012_UserStudy.pdf
49. B. Ramic-Brkic, A. Chalmers, A. Sadzak, K. Debattista, and S. Sultanic. 2013. Exploring multiple modalities for selective rendering of virtual environments. In *Proceedings of the 29th Spring Conference on Computer Graphics*. ACM, 91–98.
50. N. Ranasinghe, K-Y. Lee, and E. Y-L. Do. 2014. FunRasa: an interactive drinking platform. In *Proceedings of the 8th International Conference on Tangible, Embedded and Embodied Interaction*. ACM, 133–136.
51. N. Ranasinghe, T. N. T. Nguyen, Y. Liangkun, L-Y. Lin, D. Tolley, and E. Y-L. Do. 2017. Vocktail: A Virtual Cocktail for Pairing Digital Taste, Smell, and Color Sensations. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 1139–1147.
52. R. A. Rousi, J. M. Silvennoinen, P. M. H. Perälä, and J. P. P. Jokinen. 2017. Beyond MAYA for Game-changing Multisensory Design. In *Proceedings of the 21st International Academic Mindtrek Conference (AcademicMindtrek '17)*. ACM, New York, NY, USA, 147–153. <https://doi.org/10.1145/3131085.3131113>
53. N. Sakai, S. Imada, S. Saito, T. Kobayakawa, and Y. Deguchi. 2005. The effect of visual images on perception of odors. *Chemical Senses* 30, suppl_1 (2005), i244–i245.
54. E. B. Saleme and C. A. S. Santos. 2015. PlaySEM: A Platform for Rendering MulSeMedia Compatible with MPEG-V. In *Proceedings of the 21st Brazilian Symposium on Multimedia and the Web (WebMedia '15)*. ACM, New York, NY, USA, 145–148. <https://doi.org/10.1145/2820426.2820450>
55. E. B. Saleme, C. A. S. Santos, R. A. Falbo, G. Ghinea, and F. Andres. 2018b. Towards a Reference Ontology on Mulsemedia Systems. In *Proceedings of International Conference on Management of Emergent Digital EcoSystems, Tokyo, Japan, September, 2018 (MEDES'18)*. 8 pages. In press. <https://www.dropbox.com/s/vn39mi425a78q3t/4.pdf?dl=0>
56. E. B. Saleme, C. A. S. Santos, and G. Ghinea. 2018a. Coping with the Challenges of Delivering Multiple Sensorial Media. *IEEE MultiMedia* (2018), 11 pages. In press. <https://bura.brunel.ac.uk/handle/2438/16592>
57. H-S. Seo, A. Arshamian, K. Schemmer, I. Scheer, T. Sander, G. Ritter, and T. Hummel. 2010. Cross-modal integration between odors and abstract symbols. *Neuroscience Letters* 478, 3 (2010), 175 – 178. <https://doi.org/10.1016/j.neulet.2010.05.011>
58. S-H. Shin, K-S. Ha, H-O. Yun, and Y-S. Nam. 2016. Realistic media authoring tool based on MPEG-V international standard. In *Ubiquitous and Future Networks (ICUFN), 2016 Eighth International Conference on*. IEEE, 730–732. <https://doi.org/10.1109/ICUFN.2016.7537133>
59. J. Simner and V. Ludwig. 2009. What colour does that feel? Cross-modal correspondences from touch to colour. In *Third International Conference of Synaesthesia and Art, Granada, Spain, April*.
60. R. H. Simpson, M. Quinn, and D. P. Ausubel. 1956. Synesthesia in children: Association of colors with pure tone frequencies. *The*

- Journal of Genetic Psychology* 89, 1 (1956), 95–103.
61. F. J. Spector and D. Maurer. 2012. Making sense of scents: the colour and texture of odours. *Seeing and perceiving* 25 6 (2012), 655–77.
 62. C. Spence. 2010. The color of wine—Part 1. *The World of Fine Wine* 28 (2010), 122–129.
 63. C. Spence. 2011. Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics* 73, 4 (2011), 971–995.
 64. P. Stephen and S. Hornby. 1997. Simple statistics for library and information professionals. (1997).
 65. N. L. Streeter and T. L. White. 2011. Incongruent contextual information intrudes on short-term olfactory memory. *Chemosensory Perception* 4, 1-2 (2011), 1–8.
 66. Striner. 2018. Can Multisensory Cues in VR Help Train Pattern Recognition to Citizen Scientists? *arXiv preprint arXiv:1804.00229* (2018).
 67. Y. Sulema. 2016. Mulsemedia vs. Multimedia: State of the art and future trends. In *Systems, Signals and Image Processing (IWSSIP), 2016 International Conference on*. IEEE, 1–5.
 68. X. Sun, X. Li, L. Ji, F. Han, H. Wang, Y. Liu, Y. Chen, Z. Lou, and Z. Li. 2018. An extended research of crossmodal correspondence between color and sound in psychology and cognitive ergonomics. *PeerJ* 6 (March 2018), e4443. <https://doi.org/10.7717/peerj.4443>
 69. B. Tag, T. Goto, K. Minamizawa, R. Mannschreck, H. Fushimi, and K. Kunze. 2017. atmoSphere: Mindfulness over Haptic-Audio Cross Modal Correspondence. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. ACM, 289–292.
 70. Tanaka and A. Parkinson. 2016. Haptic wave: A cross-modal interface for visually impaired audio producers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2150–2161.
 71. Tsiros. 2017. The parallels between the study of cross-modal correspondence and the design of cross-sensory mappings. In *Proceedings of the conference on Electronic Visualisation and the Arts*. BCS Learning & Development Ltd., 175–182.
 72. S. ur Réhman, M. S. L. Khan, L. Li, and H. Li. 2014. Vibrotactile TV for immersive experience. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. IEEE, 1–4.
 73. G. Valenti and K. R. Westerterp. 2013. Optical heart rate monitoring module validation study. In *Consumer Electronics (ICCE), 2013 IEEE International Conference on*. IEEE, 195–196.
 74. M. Waltl, B. Rainer, C. Timmerer, and H. 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>
 75. M. Waltl, C. Timmerer, and H. Hellwagner. 2010. Improving the Quality of multimedia Experience through sensory effects. In *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*. 124–129. <https://doi.org/10.1109/QoMEX.2010.5517704>
 76. J. M. Yau, J. B. Olenczak, J. F. Dammann, and S. J. Bensmaia. 2009. Temporal frequency channels are linked across audition and touch. *Current Biology* 19, 7 (2009), 561–566.
 77. Z. Yuan, T. Bi, G. M. Muntean, and G. Ghinea. 2015a. Perceived Synchronization of Mulsemedia Services. *Multimedia, IEEE Transactions on PP*, 99 (2015), 1–1. <https://doi.org/10.1109/TMM.2015.2431915>
 78. Z. Yuan, S. Chen, G. Ghinea, and G-M. Muntean. 2014. User quality of experience of mulsemedia applications. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 11, 1s (2014), 15. <https://doi.org/10.1145/2661329>
 79. Z. Yuan, G. Ghinea, and G-M. Muntean. 2015b. Beyond multimedia adaptation: Quality of experience-aware multi-sensorial media delivery. *IEEE Transactions on Multimedia* 17, 1 (2015), 104–117.
 80. L. Zou, I. Tal, A. Covaci, E. Ibarrola, G. Ghinea, and G-M. Muntean. 2017. Can Multisensorial Media Improve Learner Experience?. In *Proceedings of the 8th ACM on Multimedia Systems Conference*. ACM, 315–320.
 81. Metatla, O., Maggioni, E., Cullen, C., & Obrist, M. 2019 " Like Popcorn": Crossmodal Correspondences Between Scents, 3D Shapes and Emotions in Children. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (p. 459). ACM.