



Kent Academic Repository

Mesfin, Gebremariam, Hussain, Nadia, Covaci, Alexandra and Ghinea, Gheorghita (2019) *Using Eye Tracking and Heart-Rate Activity to Examine Crossmodal Correspondences QoE in Mulsemedia*. ACM Transactions on Multimedia Computing, Communications, and Applications, 15 (2). pp. 1-22. ISSN 1551-6857.

Downloaded from

<https://kar.kent.ac.uk/77598/> The University of Kent's Academic Repository KAR

The version of record is available from

<https://doi.org/10.1145/3303080>

This document version

Author's Accepted Manuscript

DOI for this version

Licence for this version

UNSPECIFIED

Additional information

Versions of research works

Versions of Record

If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

Author Accepted Manuscripts

If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in *Title of Journal*, Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries

If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our [Take Down policy](https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies) (available from <https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies>).

Using Eye Tracking and Heart Rate Activity to Examine Crossmodal Correspondences QoE in Mulsemedia

GEBREMARIAM MESFIN, Brunel University London, United Kingdom

NADIA HUSSAIN, Brunel University London, United Kingdom

ALEXANDRA COVACI, University of Kent, United Kingdom

GHEORGHITA GHINEA, Brunel University London, United Kingdom

Different senses provide us with information of various levels of precision and enable us to construct a more precise representation of the world. Rich multisensory simulations are thus beneficial for comprehension, memory reinforcement or retention of information. Crossmodal mappings refer to the systematic associations often made between different sensory modalities (e.g., high pitch is matched with angular shapes) and govern multisensory processing. A great deal of research effort has been put in exploring crossmodal correspondences in the field of cognitive science. However, the possibilities they open in the digital world have been relatively unexplored. Mulsemedia - multiple sensorial media - provides a highly immersive experience to the users and enhances their Quality of Experience (QoE) in the digital world. Thus, we consider that studying the plasticity and the effects of crossmodal correspondences in a mulsemedia setup can bring interesting insights about improving the human computer dialogue and experience. In our experiments, we exposed users to videos with certain visual dimensions (brightness, color and shape) and we investigated if the pairing with a crossmodal matching sound (high and low pitch) and the corresponding auto-generated vibrotactile effects (produced by a haptic vest) lead to an enhanced QoE. For this, we captured the eye gaze and the heart rate of users while experiencing mulsemedia and we asked them to fill in a set of questions targeting their enjoyment and perception at the end of the experiment. Results showed differences in eye gaze patterns and heart rate between the experimental and the control group indicating changes in participants' engagement when videos were accompanied by matching crossmodal sounds (this effect was the strongest for the video displaying angular shapes and high pitch audio) and transitively generated crossmodal vibrotactile effects.

CCS Concepts: • **Human-centered computing** → **HCI design and evaluation methods;Interaction paradigms;Interaction techniques;Empirical studies in HCI;**

Additional Key Words and Phrases: Mulsemedia, video, audio, haptic, crossmodal correspondence, quality of experience, gaze tracking, heart rate variability

ACM Reference Format:

Gebremariam Mesfin, Nadia Hussain, Alexandra Covaci, and Gheorghita Ghinea. 2018. Using Eye Tracking and Heart Rate Activity to Examine Crossmodal Correspondences QoE in Mulsemedia. *ACM Trans. Multimedia Comput. Commun. Appl.* x, x, Article x (January 2018), 21 pages. <https://doi.org/0000001.0000001>

This paper was funded by the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 688503.

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; Alexandra Covaci, University of Kent, United Kingdom, a.covaci@kent.ac.uk; Gheorghita Ghinea, Brunel University London, United Kingdom, george.ghinea@brunel.ac.uk.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2018 Association for Computing Machinery.

Manuscript submitted to ACM

Manuscript submitted to ACM

1 INTRODUCTION

We rarely experience senses in isolation. Real world events produce different sensory signals that build our perceptual experience as a coherent form through the integration of information from various sensory channels. Multisensory integration occurs at the intersection between two or more sensory modalities and is facilitated by space and time coincidence constraints and enhanced by semantic and synesthetic (crossmodal) correspondences. In the digital world, research results from the (traditionally) non-digital field of cognitive science regarding the crossmodal mappings might benefit the design and evaluation of user interfaces and information display.

Mulsemmedia enriches traditional audio video content with new media types such as olfactory, haptic and gustatory, making possible the inclusion of layered sensory stimulation and interaction between multiple sensory channels. Mulsemmedia brings new opportunities for the development of immersive technologies and also opens new perspectives in real world applications including education, communication, advertising and medicine [Ghinea et al. 2014; Sulema 2016].

In this emerging research field, there are various explorations on the practicality and possibility of incorporating different media types into applications focused mainly on the digital representation of sensing, storage and display, and less on its impact on Quality of Experience [Cingel and Piper 2017; Ghinea and Ademoye 2012a,b; Murray et al. 2017]. QoE refers to the "degree of delight or annoyance of applications or services" [Brunnström et al. 2013].

In [Murray et al. 2017], the authors emphasize that mulsemmedia has a major role in increasing the level of immersion and the QoE. However, the existing literature has limitation in showing the impact of crossmodally mapped audio (based on selected features of a video) and its (transitively) auto-generated haptic effect on QoE. Moreover, whilst correspondences have been explored in the non-digital world, not the same can be said as far as the case of digital mulsemmedia is concerned.

Given the mechanisms behind multisensory integration, we hypothesize that an important factor in enhancing QoE could be the crossmodal mappings between the various media dimensions, as played out in the digital world. Thus, in this paper, we focus on determining how selected crossmodal associations impact the users' perceived QoE. That is, we intend to explore the crossmodal correspondence between visual features of selected videos and the corresponding audio pitch through auto-generated vibrotactile feedback, and the overall impact on QoE.

For our study we chose to combine different visual dimensions (brightness, color and shape) with crossmodal matched audio dimensions (high and low pitch) and with auto-generated (from audio) vibrotactile feedback. Since the vibrotactile feedback is automatically transformed from the audio associated with the audio-visual content, our method bypasses the translation from visual to haptic modality while keeping the correspondences valid through transitivity. For measuring QoE, we use both objective metrics (heart rate - also used in [Tortell et al. 2007] and eye gaze tracking [Gulliver and Ghinea 2004]) and subjective metrics (questions with answers on a Likert scale), as recommended in [Murray et al. 2017], where the authors highlight issues in evaluating QoE in mulsemmedia.

The remainder of the paper is organized as follows: Section 2 provides an overview of the state of the art in areas related to mulsemmedia, QoE and crossmodal correspondences. Section 3 presents the methodology we used in running the experiments. Section 4 illustrates our results, while Section 5 concludes the paper.

2 RELATED WORK

The perception of individual senses has been studied extensively both in the cognitive and in the digital world. Crossmodal correspondences have been addressed mainly in psychology with interesting results but also with still

105 unanswered questions. This phenomena has been also studied in the computer graphics field in the context of increasing
106 the rendering performance and achieving more realistic virtual environments. Since these results were promising, we
107 hypothesize that considering crossmodal mappings in the design of mulsemmedia systems also has to be investigated
108 since it could lead to more user effective experiences.
109

113 2.1 Auditory-visual crossmodal correspondences research in psychology

114
115 Experiencing a stimulus in a sensory modality is often associated with experiences in another sensory dimension
116 (e.g., pitch in audition and brightness in vision). While the focus in our paper is on the nature of the crossmodal
117 correspondences that exist between auditory and visual stimuli, crossmodal associations were documented between
118 many pairs of sensory modalities such as: vision and touch [Simmer and Ludwig 2009], audition and touch [Yau et al.
119 2009], flavors and sounds [Crisinel and Spence 2009], smell and vision [Seo et al. 2010; Spence 2010], flavors and vision
120 [Gal et al. 2007] - all in a non-digital context.
121

122 As demonstrated by researchers in different experimental approaches, nonarbitrary crossmodal correspondences
123 exist between a variety of auditory and visual stimuli. These mappings have been documented between simple stimulus
124 dimensions (loudness and brightness) and between more complex stimuli (shapes/images, words). According to Marks
125 [Marks 1987], higher pitch and louder sounds are associated with lighter colors. The correspondence between frequency
126 of music and lightness of a colored object was investigated in an eye-tracked study in [Hagtvedt and Brasel 2016].
127 Results showed that visual attention was guided towards light-colored objects under the influence of high frequency
128 sounds. This gives a clear example on how the soundtrack frequency can be used to manipulate the attention of the user
129 to certain things, with promising applications in marketing strategies. In addition, there is evidence that sound is also
130 matched to other visual attributes like shapes: high pitched tones are matched with angular shapes, while low-pitch
131 sounds are matched with rounder shapes [Marks 1987].
132

133 It is important to emphasize that this interaction between audition and vision is bidirectional: audition influences
134 vision and vision influences auditory perception. The modality that is more appropriate for a particular task will
135 dominate the perception in that particular task - vision dominates perception in the spatial domain (e.g., ventriloquism
136 effect) while in the temporal domain, audition is dominant [Welch and Warren 1980]. There are different types of effects
137 audio can have on the visual perception: (i) driving effect - when presented simultaneously, sound drives vision in the
138 temporal domain [Welch et al. 1986], (ii) illusory flash - an illusory flash can be induced by a sound beep [Shams et al.
139 2000].
140

141 The bidirectionality of crossmodal correspondences can also mean that these associations are transitive and indeed,
142 transitivity is one of the properties that differentiate crossmodal correspondences from synaesthesia. If we know how
143 a visual feature (e.g., shape) is related to audio pitch (e.g., round objects make low pitched sounds) and how audio
144 pitch is related to density, then we should be able to predict how shape is related to density. Transitivity has not been
145 investigated thoroughly, but the multidimensionality of the precepts at stake seems to indicate that it should not be
146 expected in every case. An example where transitivity does not stand is between loudness and size (where louder
147 corresponds to larger) and the crossmodal correspondence between pitch and size (where lower pitch corresponds
148 to larger). Although transitivity would indicate that louder corresponds to lower pitch, research outputs showed the
149 opposite [Spence 2011].
150
151
152
153
154
155
156

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

Despite the fact that crossmodal correspondences between visual and auditory media show an interesting potential, the work on this outside the field of cognitive sciences is limited. The influence of auditory stimuli on visual perception was investigated in [Mastoropoulou 2007; Mastoropoulou et al. 2005], where the authors showed that when only sound emitting objects are rendered in high quality and the remainder of the scene in much lower quality, the visual quality is not affected. Selective rendering based on crossmodality principles were analyzed also in [Brkic et al. 2009], this time the authors focusing on the mapping between smell and visual. The study showed that the smell of fresh cut grass can distract viewers from correctly identifying the animation quality (flyover of a grass terrain). In [Hulusić et al. 2009], the authors focused on the effect of beat rate in static scenes and showed that lower beat rates influence the perception of low frame rates. Another interesting study of the same authors investigated how camera movement speed and the sounds affect the smoothness of the animation [Hulusic et al. 2010]. Their results showed that walking animations are perceived smoother and that adding movement-related sound leads to an improvement of quality. In [Ramic-Brkic et al. 2013], the authors run experiments to determine how the presence of different modalities (auditory, smell, and ambient temperature) affects the viewer's ability to perceive the quality of the graphics. The perceived rendering quality was significantly influenced by strong perfume, high temperature and audio noise [Ademoye et al. 2016]. However, [Ademoye et al. 2016] merely describes the ability of audio on masking synchronization skews in olfaction-enhanced multimedia presentations and does not consider cross-modality.

Crossmodal correspondences have found applications also outside the selective rendering theme. In [Tanaka and Parkinson 2016], the authors present Haptic Wave, a device that allows the crossmodal mapping between digital audio and the haptic domain dedicated to audio producers with visual impairments. The influence of auditory on haptic was exploited in [Koizumi et al. 2011] for designing a device (Chewing Jockey) that augments the food texture using sound. The study in [Slocombe et al. 2016] also showed that taste components can be influenced by the tactile quality of food. This solution is meant to improve the quality of eating for dentures users. Other studies utilized crossmodal perception to change the taste of food [Ranasinghe et al. 2017] or to project more abstract color/shape/music combinations over a food product [Huisman et al. 2016]. A solution for meditation support built on top of the same principles guides the users through the breathing techniques with the help of a haptic ball [Tag et al. 2017].

These applications of crossmodality concepts outside the field of psychology show that these mappings have a promising potential in designing interfaces and displays that tap into users' mental models [Tsiros 2017]. We think that testing the effectiveness of crossmodal mappings in other contexts might bring interesting insights that could allow to significantly improve the human-computer dialogue.

2.3 Mulsemedia and QoE

The last decade of research has brought an increasing interest in developing multimedia applications that incorporate media components outside the traditional audio-video (AV) content [Ghinea et al. 2014]. The research community is continuously extending the list of new media that can be added on top of discrete and continuous media like text, animation, graphics, audio, and video. These new media target different other senses such as touch [Eid et al. 2008], smell [Ghinea and Ademoye 2012a] or taste [Ranasinghe et al. 2014, 2017] with the aim to increase the user's QoE and to explore novel methods for interaction [Murray et al. 2017].

Mulsemmedia systems are created as a result of a production-distribution-rendering workflow that can be seen as an extension of the process described in [Danieau et al. 2013] for adding haptic effects to AV. Production deals with how different sensory effects can be created or generated in synchronization with the AV content. This can be done through digital capturing and processing of data acquired from sensors, automatic extraction from audio or AV content or manual authoring. Once the mulsemmedia effects are captured, they are encoded for transport, processed and emitted for distribution to providers, distributed to the end-users and then decoded and rendered by different devices and experienced by the end users. In designing effective mulsemmedia experiences it is important to identify an optimal and consistent way to deliver different sensory content. In [Murray et al. 2017], it is stated that a motivation for pursuing more mulsemmedia components is to increase the level of user immersion and/or QoE.

The evaluation of mulsemmedia systems through smell, taste, sight, touch, temperature, balance perception can contribute to a high degree of qualitative differentiation when it comes to factors like QoE. In [Brunnström et al. 2013], the QoE is described as the degree of delight or annoyance of the user of an application or service, and traditionally, its aspects can be assessed using either perception-based (subjective) or instrumental (objective) methods. The most valid way to assess the QoE is represented by subjective methodologies that require human evaluators to gather QoE-related information for different multimedia stimuli. Objective methods are low-cost and a quicker alternative because they use an algorithm or an instrument. QoE assesses the intentions, satisfaction and emotions triggered by a particular product or service. Thus, when it comes to evaluate the QoE, taking the emotional aspects into consideration could bring interesting insights. Recently, physiological measures (e.g., heart rate, skin conductance, skin temperature) have been shown to be useful in affective state monitoring, particularly within a multimodal setup [Koelstra et al. 2012]. In [Egan et al. 2016], the authors presented objective and subjective QoE evaluations of virtual environments. Correlations between objective metrics (heart rate and electrodermal activity) and user subjective QoE measured via a post-test questionnaire were analyzed. Elevated values for heart rate and electrodermal activity were correlated with physiological arousal - one of the factors associated with user QoE. The potential and benefits of using these objective metrics as indicators of user QoE for immersive experiences were shown also for augmented reality applications in [Keighrey et al. 2017].

Mulsemmedia seems to contribute to the QoE directly, by fulfilling users' expectations and increasing the utility/enjoyment of the application or service, but also indirectly. For example, studies in [Ademoye et al. 2016; Yuan et al. 2014, 2015] have pointed out that multimedia sequences integrated with mulsemmedia effects can partly mask an AV sequence's decreased quality as well as synchronization skews, thus enhancing the user's perceived QoE. Moreover, designing technology that engages multiple senses has the potential to facilitate memory [Ademoye and Ghinea 2013], to improve virtual realism, to more easily transfer information between physical and digital environments [Zou et al. 2017] and to contribute to pattern recognition [Striner 2018]. For example, a study in [Ranasinghe et al. 2018] presents the use of multisensory virtual reality (VR) aiming at augmenting the sense of presence in a virtual environment. The authors indicated that the addition of any singular modality into the conventional audio-visual media improves sense of presence and providing a combination of modalities (e.g., olfaction, wind, and thermal) improves it further.

Since increasing sensory modalities has the potential to improve the overall user experience, it is important to explore the possibilities brought by substituting a sense with a combination of others senses or what happens when a sensory stimuli is stronger than the others. Crossmodal correspondences were rarely considered in the design of mulsemmedia systems although our perceptual experience is strongly influenced by them. Crossmodal interactions present advantages because they can help to overcome a specific sensory deprivation or situational impairment (e.g., auditory and tactile senses in darkness) [Hoggan and Brewster 2006]. Based on these promising results, we hypothesize that mulsemmedia

can be a context with potential to enrich the knowledge about crossmodal correspondences, that have been investigated until now mostly in setups based on image and text samples. Moreover, crossmodal correspondences could be used in the development of guidelines for the effective design of 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 [ur Réhman et al. 2014].

3.1 Participants

24 participants (15 males, 9 females) took part in the experiment and were randomly assigned to equal-sized experimental and control groups. Users were aged from 18 to 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 running on a Windows 10 Laptop with 8GB RAM powered by an IntelCore i5 processor. The viewing screen was placed between 45-75 cm from the eyes of the participants, as this was the recommended distance for Eye Tribe calibration¹. We chose to use the EyeTribe eye tracker because this was demonstrated to be accurate enough in studies on gaze points and fixations [Dalmaijer 2014].

Participants sat in a chair without armrests facing the screen. All participants wore i-shine² headphones, a vibrotactile KOR-FX³ gaming vest, and a Mio Link heart rate wristband⁴. 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. Users of this type of devices obtain additional information about environmental factors while becoming more immersed/involved in their experience [Munster et al. 2015]. A snapshot of the experimental setup is shown in Fig. 1.

Two software artifacts are also developed (a media player, and eye-tracking data recorder) with Java. The media player is built using the JavaFX⁵ MediaPlayer package which we used to play the sample videos (in synchrony with the altered audio) with additional functionalities such as play-time recording, and on-screen questions rendering and response management. The eye-tracking data recorder, on the other hand, is developed using the TheEyeTribe Java

¹<http://theeyetribe.com/dev.theeyetribe.com/dev.theeyetribe.com/start/index.html>

²<https://www.ishine-trade.com/Headphones-Earphones>

³<http://korfx.com/products>

⁴<https://www.mioglobal.com/>

⁵<https://docs.oracle.com/javafx/2/api/javafx/scene/media/MediaPlayer.html>

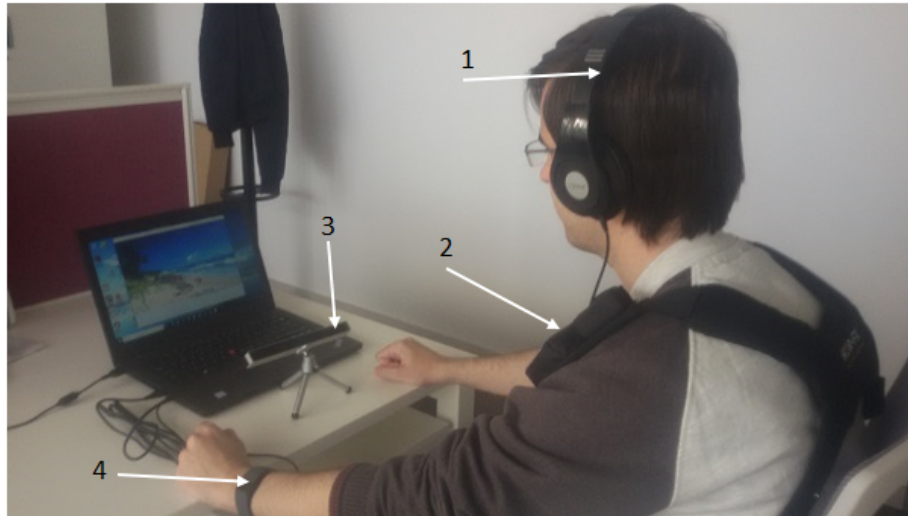


Fig. 1. Experimental setup. The users were wearing: (1) a pair of i-shine headphones, (2) the KOR-FX haptic vest, their eye gaze was captured with (3) the EyeTribe eye tracker, while their heart rate was measured with (4) Mio Link.

package which filters and records separate data about eye-gaze coordinates and recording time for each of the video samples in a single file.

3.3 Audiovisual 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 accompanying sound was modified to respect the principles of auditory-visual cross-modal correspondences as demonstrated in the literature. The video with dominant yellow images (V1) was watched accompanied by high pitch sounds, while the one dominantly blue (V2) by low pitch sounds [Simpson et al. 1956; Sun et al. 2018]. Then, when brightness was considered the dominant visual characteristic, low pitch sounds were used for the video with low level of brightness (V3), while the one with high brightness values contained high pitch (V4) sounds, based on [Marks 1974]. Finally, the video displaying angular shapes (V5) was matched with high pitch sounds and the one with round shapes (V6) with low pitch sounds [Spence 2011].

All videos in our experiment were 120 seconds long. For the experimental group, the audio was adjusted to a frequency of 328Hz (high pitch condition) and 41Hz (low pitch condition). Participants in the control group experienced the videos with their original soundtrack.

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.

Conditions. There were two conditions related to the type of audio.

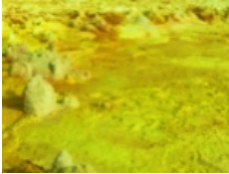

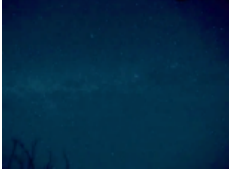


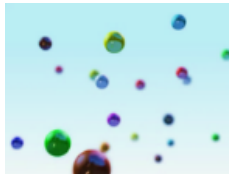
Video Snapshot	Description	Video Snapshot	Description
	V1: Theme: Yellow sulphur springs (Danakil Desert, Ethiopia) Visual dominant: Yellow color EG audio: High pitch CG audio: Smooth sound		V2: Theme: Blue waves Visual dominant: Blue color EG audio: Low pitch CG audio: Smooth sound
	V3: Theme: Solar eclipse - sky turns dark Visual dominant: Low brightness EG audio: Low pitch CG audio: Smooth sound		V4: Theme: Bright sun shining upon the Arctic Visual dominant: High brightness EG audio: High pitch CG audio: Smooth sound
	V5: Theme: Fly over angular skyscrapers Visual dominant: Angular shape EG audio: High pitch CG audio: Smooth sound		V6: Theme: Bouncing balls Visual dominant: Round shape EG audio: Low pitch CG audio: Smooth sound

Table 1. Snapshots from the videos used during the experiment and their audio visual attributes.

- The experimental condition (carried out by the experimental group EG) corresponds to the use of altered audio (original audio completely replaced and audio pitch modified) so that it matches certain dominant visual features.
- The control condition (carried out by the control group CG) corresponds to the use of non-altered audio (smooth original sound).

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 points situated on a 3x3 grid (Figure 2). The illustration in Figure 2 shows the points as circular targets displayed on the screen on a blank background. Once the calibration process is completed, the system begins providing (x, y) screen coordinates of the user's gaze point.

The viewing screen was placed 50 cm from the eyes of the participants which is within the recommended range of the Eye Tribe calibration⁶.

Participants were randomly divided in two groups of 12 each and watched the six videos in a random order for both control and experimental groups. 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:

⁶<http://theyetribe.com/dev.theyetribe.com/dev.theyetribe.com/start/index.html>

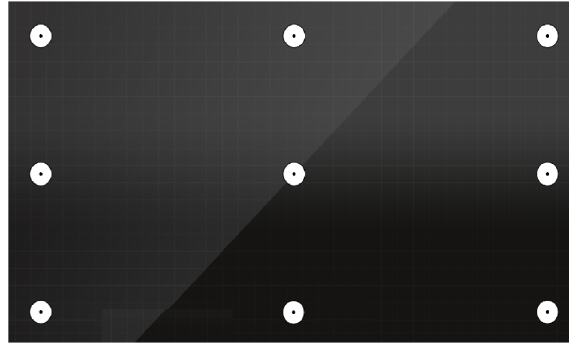


Fig. 2. A Screen shot of Eye-gaze Calibration

- 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 [Valenti and Westerterp 2013]. 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.

Participants also completed a subjective questionnaire 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.

- Q1: I enjoyed watching the video clip whilst wearing a haptic vest.
- Q2: The haptic vest effects were relevant to the video clip I was watching.
- Q3: The vibration was annoying.
- Q4: The haptic vest effects enhanced the sense of reality whilst watching the video clip.
- Q5: The haptic vest effects enhanced my viewing experience.
- Q6: The haptic vest effects were necessary when watching a video clip.
- Q7: The temperature of the haptic vest came across warm.
- Q8: The vibration was distracting.
- Q9: I enjoyed watching the video clip.

4 RESULTS AND DISCUSSION

This section presents analysis and discussion of results of the data collected via the eye-tracker, heart-rate monitor, and on-screen QoE questionnaire. 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 [Stephen and Hornby 1997], 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 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 control and experimental groups 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 experimental and control groups across all cells, as shown in equation (1).

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

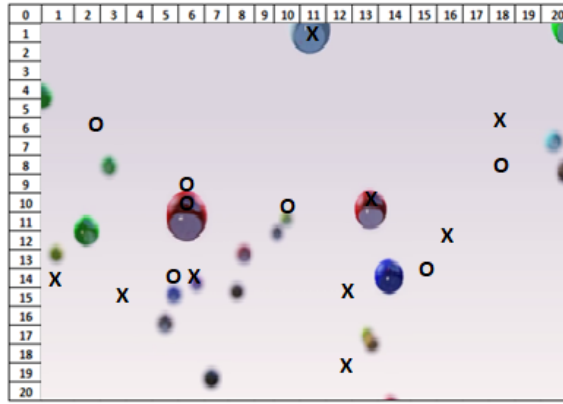


Fig. 3. Points where the participants gazed at the 10th frame of video V6 (X=Experimental Group, O=Control Group)

In this regard, the minimum and maximum eye gaze difference count between the experimental and control groups are $\text{Min } \Delta = 0$ and $\text{Max } \Delta = 24$, respectively. For example, Figure 3 shows the eye gaze count at the 10th frame of video 6 observed from participants in both control and experimental groups.

The eye gaze difference count between the experimental and control groups for all frames in each video is depicted in Figure 4 and shows non-zero values ranging between 2 and 10 at each video frame. That is, between 2 to 10 participants starred at different cells on the viewing area.

As can be seen from Figure 4, participants' viewing patterns on each of the video clips show varying axes and amplitudes of eye gaze count differences. Accordingly, 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 2. This reveals that there are statistically significant differences in eye gaze between the experimental and control groups for all the six videos ($p < 0.05$). As the only difference between the groups is the accompanying audio soundtrack (the control group had the original soundtrack, whilst the experimental group had a mapped high/low pitch sound), we deduce that the difference in eye gaze count is due to the difference in the experienced audio between the groups.

Video	Sig. (2-tailed)	Mean difference	95% Confidence Interval	
			Lower	Upper
V1	0.000	-5.779	-5.82	-5.73
V2	0.000	-5.572	-5.62	-5.53
V3	0.000	-6.594	-6.64	-6.55
V4	0.000	-5.903	-5.95	-5.86
V5	0.000	-5.968	6.00	-5.93
V6	0.000	-5.579	-5.62	-5.54

Table 2. One sample t-test of eye gaze difference count

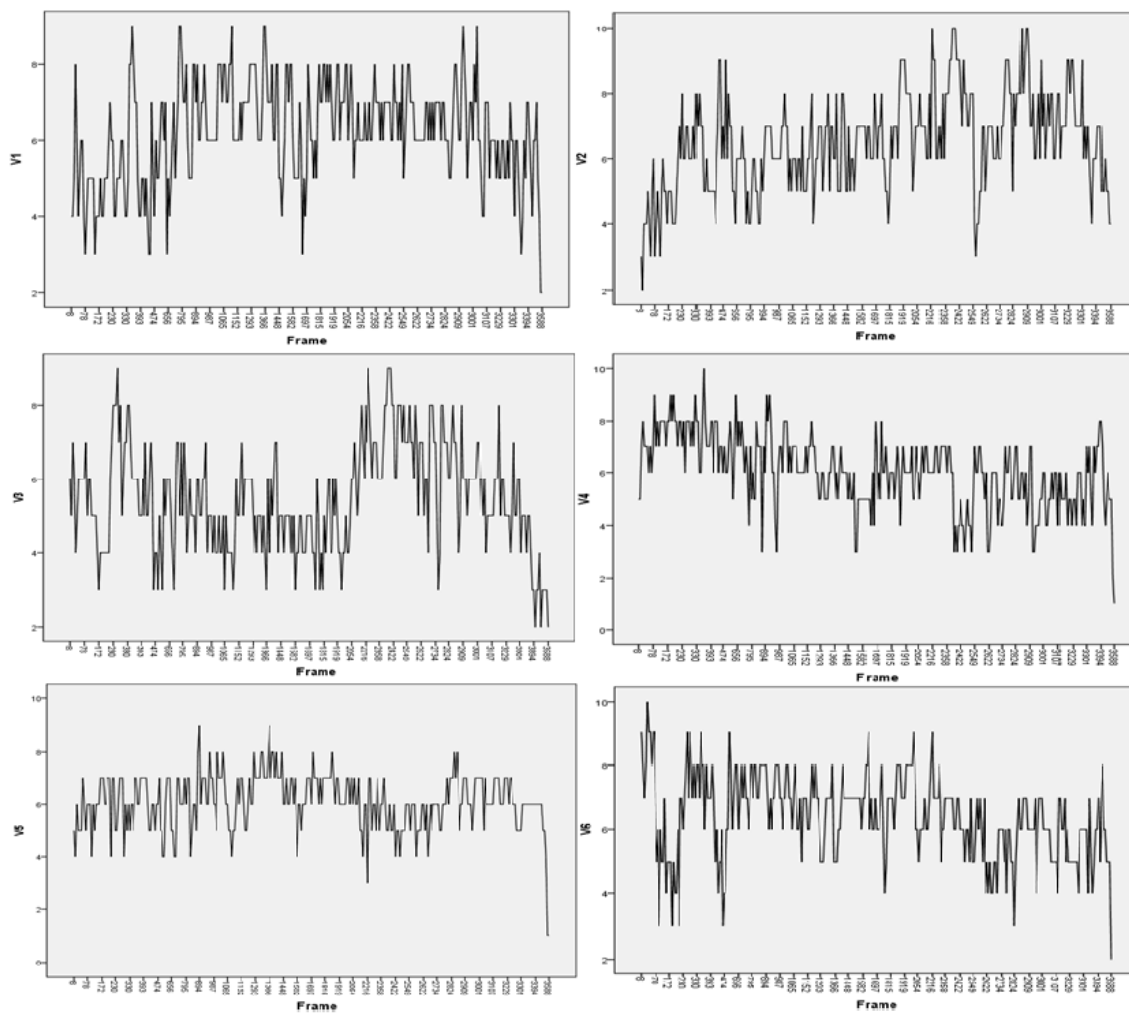


Fig. 4. Eye gaze difference count across video clips

4.2 Analysis of heart rate data

As outlined before, the physiological metric employed in this study was the heart rate captured by the Mio Link. Heart rate data 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 control group varied between 50bpm and 100bpm whilst, for the experimental group, these ranged between 80bpm and 110bpm, with the means for each video illustrated in Figure 5.

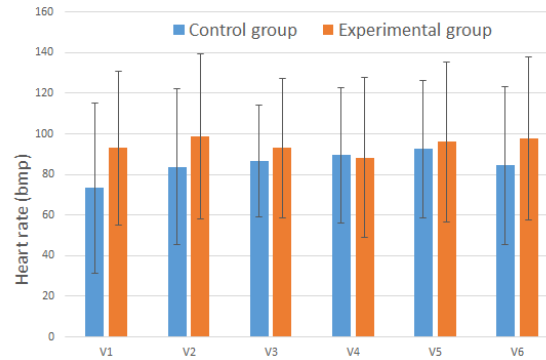


Fig. 5. Average heart rate data for all video clips

In Figure 6 we present the mean heart rate gathered every second for each of the six videos in both control and experimental group. We observe a tendency for a higher heart rate in the experimental group for the whole duration of the videos (except for V4).

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 3. The results in Table 3 demonstrate a statistically significant difference between the heart rates of the two groups for all the videos, except V4 that depicted a high brightness scene.

This indicates that the two groups experienced a different mood in the two setups: (i) the one using crossmodally matching sound (experimental group) and (ii) the one where the sound did not correspond to the visual dominant features (control group). We remind the reader, that the sound served as an input also for the haptic vest which automatically converts it into the vibrotactile feedback (Section 3.2). The most significant differences in heart rate appear for V1, V2 and V6 (as presented in Table 3). This shows that something in the user's mood changes when she/he experiences content where high pitch sounds are matched with yellow color and low pitch sounds are matched with blue or round objects. Even if smaller, the differences between the two groups are significant also for V3 and V5. This indicates that angular objects matched with high pitch sounds and dark images matched with low pitch sounds produce also a change in the affective state of the user.

4.3 Analysis of self-reported QoE

Participants self-reported the QoE by answering a series of 9 Likert scale questions, as detailed in Section 3.4. Before analyzing the data, we converted the scores of each negatively-phrased question to the equivalent score associated with a positively-phrased counterpart.

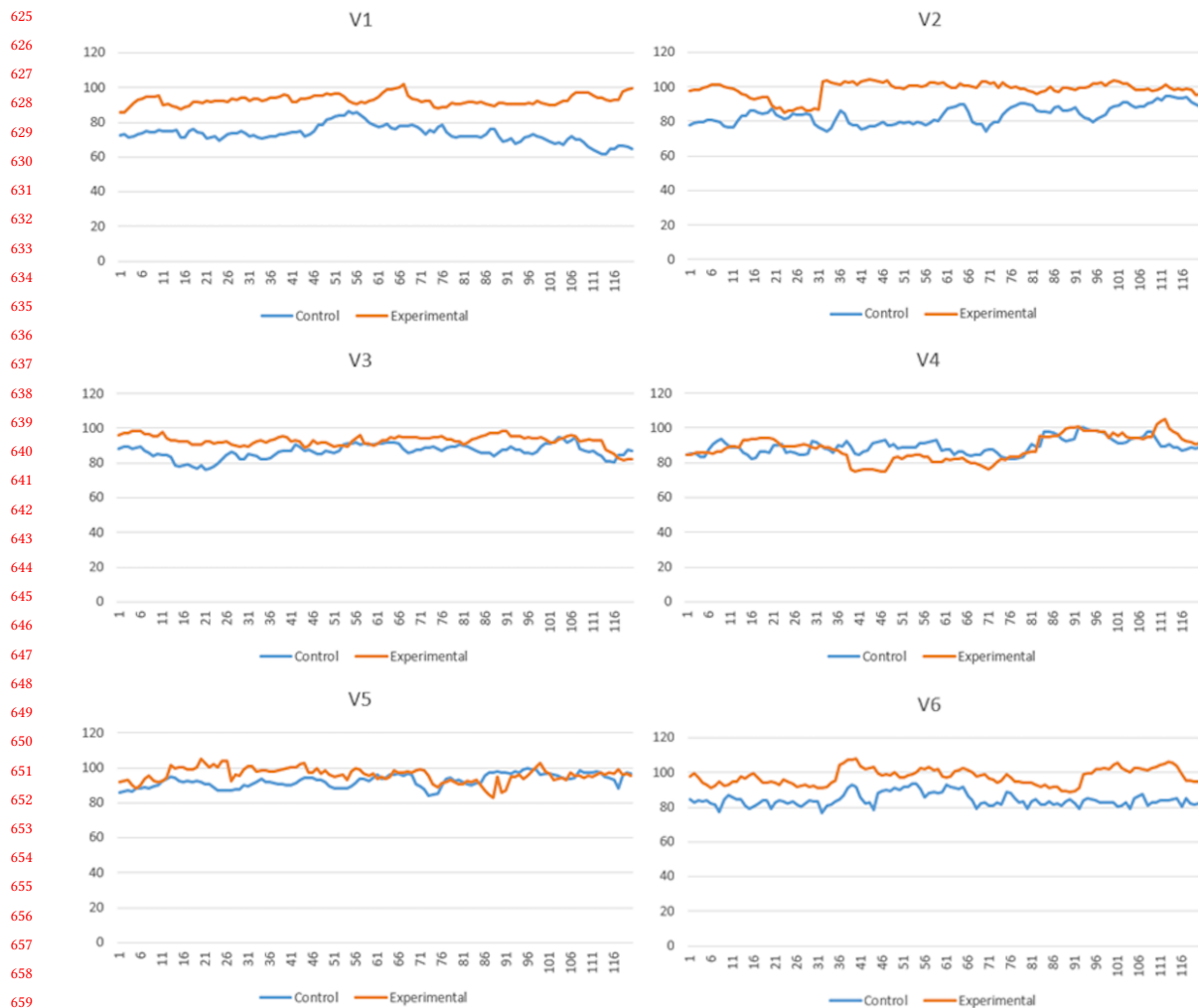


Fig. 6. Average heart rate data (bpm) of the participants for each video

In a first instance, we performed a two way ANOVA with group type and video type as independent variables and the responses to the 9 self-reported QoE questions as the dependent variables, the results of which are presented in Table 4. As it can be seen in Table 4, the interaction of the independent variables (Group*Video) has a statistically significant effect for all questions except Q9 (*I enjoyed watching the video clip*) on the self-reported QoE (dependent variable). Since the interaction was statistically significant we undertook simple main effects analysis in respect of Q1-Q8 and Post Hoc tests for Q9, respectively. The results of these investigations are presented next.

Q1: I enjoyed watching the video clip whilst wearing a haptic vest. The results of simple main effect analysis of the impact of the independent variables on Q1 are shown in Table 5. These revealed that for V3, the high/low pitched audio (and associated haptic effect) led to statistically significant higher scores in the experimental group participants compared to their control group counterparts ($F(1,132) = 10.154, p < .05$). However, for V2 ($F(1,132) = 6.268, p < .05$) and

Video	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval	
				Lower	Upper
V1	0.000	-19.616	1.482	-22.521	-16.711
V2	0.000	-14.877	1.487	-17.794	-11.961
V3	0.000	-6.312	1.160	-8.584	-4.040
V4	0.354	1.260	1.362	-1.405	-3.932
V5	0.013	-3.403	1.369	-6.086	-0.719
V6	0.000	-13.370	1.468	-16.248	-10.943

Table 3. Independent samples t-test of heart rate data

V6 ($F(1,132) = 10.154, p < .05$), the situation is reversed, with control group participants reporting statistically significant higher scores in respect of their enjoyment of watching the respective videos whilst wearing haptic vests. However, in the case of V1, V4, and V5, differences in participant scores were not significant between the two groups.

Q2: The haptic vest effects were relevant to the video clip I was watching. The results of simple main effect analysis of the independent variables in respect of Q2 are shown in Table 6. These showed that for V3, there was a statistically significant difference between experimental and control groups, with the former reporting higher scores than the latter ($F(1,132) = 7.797, p < .05$). This situation is reversed for V6 ($F(1,132) = 6.723, p < .05$), where nonetheless differences remain statistically significant. In the case of V1, V2, V4, and V5, however, differences in participants' scores were not statistically significant between experimental and control groups.

Source	Dep. Var.	df	F	Sig.	Source	Dep. Var.	df	F	Sig.	Source	Dep. Var.	df	F	Sig.
Group	Q1	1	4.625	0.033	Video	Q1	5	2.826	0.019	Group*Video	Q1	5	5.218	0.000
	Q2	1	0.004	0.953		Q2	5	4.019	0.002		Q2	5	3.363	0.007
	Q3	1	0.238	0.627		Q3	5	4.493	0.001		Q3	5	9.518	0.000
	Q4	1	0.475	0.492		Q4	5	2.659	0.025		Q4	5	5.765	0.000
	Q5	1	4.414	0.038		Q5	5	1.360	0.243		Q5	5	6.478	0.000
	Q6	1	3.356	0.069		Q6	5	2.402	0.040		Q6	5	5.471	0.000
	Q7	1	2.809	0.096		Q7	5	6.908	0.000		Q7	5	2.966	0.014
	Q8	1	0.001	0.970		Q8	5	3.643	0.004		Q8	5	9.877	0.000
	Q9	1	0.406	0.525		Q9	5	2.844	0.018		Q9	5	1.504	0.193

Table 4. ANOVA multivariate test result for each question

Video	df	F	Sig.	Mean Difference (Experimental-Control)	Video	df	F	Sig.	Mean Difference (Experimental-Control)
V1	1	0.466	0.496	-0.250	V1	1	0.040	0.842	-0.083
V2	1	6.268	0.014	-0.917	V2	1	0.159	0.691	-0.167
V3	1	10.154	0.002	1.167	V3	1	7.797	0.006	1.167
V4	1	1.865	0.174	-0.500	V4	1	0.000	1.000	0.000
V5	1	0.207	0.650	-0.167	V5	1	0.358	0.551	0.250
V6	1	10.154	0.002	-1.167	V6	1	6.723	0.011	-1.083

Table 5. Simple main effects analysis (Q1)

Table 6. Simple main effects analysis (Q2)

Video	df	F	Sig.	Mean Difference (Experimental- Control)
V1	1	8.790	0.004	1.083
V2	1	0.052	0.820	0.083
V3	1	22.936	0.000	1.750
V4	1	1.872	0.174	-0.500
V5	1	5.201	0.024	-0.833
V6	1	10.194	0.002	-1.167

Table 7. Simple main effects analysis (Q3)

Video	df	F	Sig.	Mean Difference (Experimental- Control)
V1	1	3.725	0.056	-0.750
V2	1	4.599	0.034	-0.833
V3	1	12.042	0.000	1.417
V4	1	0.184	0.669	-0.167
V5	1	2.254	0.136	0.583
V6	1	4.599	0.034	-0.833

Table 8. Simple main effects analysis (Q4)

Q3: The vibration was annoying. The results of simple main effect analysis of the independent variables in respect of Q3 are shown in Table 7. These showed that for V1 ($F(1,132) = 8.790, p < .05$) and V3 ($F(1,132) = 22.936, p < .05$), there were statistically significant differences between experimental and control groups as regards users deeming the vibrations of the haptic vest annoying. Given our mirrored coding of negative statements when analyzing the QoE questionnaire, these showed that haptic vest vibrations were found to be less annoying by experimental group participants for V1 and V3. Significant differences were also found for V2 ($F(1,132) = 5.201, p < .05$) and V6 ($F(1,132) = 10.194, p < .05$). However, in these cases, the situation is reversed, with control group participants finding the haptic vest vibrations to be less annoying than their experimental group counterparts. In the case of V2 and V4, there were no statistically significant differences between the two groups.

Q4: The haptic vest effects enhanced the sense of reality whilst watching the video clip. The results of simple main effect analysis of the independent variables in respect of Q4 are shown in Table 8. These highlighted that there were statistically significant differences for V3 ($F(1,132) = 13.292, p < .05$), V2 ($F(1,132) = 4.599, p < .05$) and V6 ($F(1,132) = 4.599, p < .05$) between experimental and control groups, in respect of the haptic effects experienced enhancing the sense of realism. In the cases of V3 users in the experimental group gave higher scores than their control group counterparts, whilst for V2 and V6 the situation is reversed. Lastly, in the cases of V1, V4, and V5, there were no significant differences in scoring between the two groups.

Q5: The haptic vest effects enhanced my viewing experience. The results of simple main effect analysis of the independent variables in respect of Q5 are shown in Table 9. These yield statistically significant differences between experimental and control groups for V3 ($F(1,132) = 16.720, p < .05$), V1 ($F(1,132) = 6.669, p < .05$) and V6 ($F(1,132) = 5.604, p < .05$). In the case of V3 it was experimental groups users who gave higher scores than control group participants in respect of haptic vests enhancing the viewing experience, whilst for V1 and V6 the situation is reversed. No significant differences were found between the two groups in the case of V2, V4, and V5.

Q6: The haptic vest effects were necessary when watching a video clip. The results of simple main effect analysis of the independent variables in respect of Q6 are shown in Table 10. These show that for V3 ($F(1,132) = 9.110, p < .05$), V1 ($F(1,132) = 7.936, p < .05$), V2 ($F(1,132) = 4.049 < .05$) and V6 ($F(1,132) = 9.110, p < .05$), there were statistically significant differences between experimental and control group scores regarding haptic effects being necessary whilst watching the respective video clips. Whilst for V1 it was experimental group users who felt more strongly in this respect, the converse is true for V1, V2 and V6. No significant differences were found for V4 and V5.

Q7: The temperature of the haptic vest came across warm. The results of simple main effect analysis of the independent variables in respect of Q7 are shown in Table 11. The analysis only revealed statistically significant scores

781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832

Video	df	F	Sig.	Mean Difference (Experimental- Control)
V1	1	6.669	0.011	-1.000
V2	1	2.964	0.087	-0.667
V3	1	16.720	0.000	1.583
V4	1	3.752	0.055	-0.750
V5	1	0.185	0.668	-0.167
V6	1	5.604	0.019	-0.917

Table 9. Simple main effects analysis (Q5)

Video	df	F	Sig.	Mean Difference (Experimental- Control)
V1	1	7.936	0.006	-1.167
V2	1	4.049	0.046	-0.833
V3	1	9.110	0.003	1.250
V4	1	0.040	0.841	-0.083
V5	1	0.364	0.547	0.250
V6	1	9.110	0.003	-1.250

Table 10. Simple main effects analysis (Q6)

for V6 ($F(1,132) = 14.341, p < .05$) with experimental group perceiving the temperature generated by the haptic vest as being warmer than users from the control group. No significant differences were found between the two groups in respect of the remaining videos.

Video	df	F	Sig.	Mean Difference (Experimental- Control)
V1	1	0.794	0.375	-0.333
V2	1	0.794	0.375	-0.333
V3	1	0.050	0.824	0.083
V4	1	0.050	0.824	-0.083
V5	1	1.241	0.267	0.417
V6	1	14.341	0.000	-1.417

Table 11. Simple main effects analysis (Q7)

Video	df	F	Sig.	Mean Difference (Experimental- Control)
V1	1	8.367	0.004	1.167
V2	1	0.043	0.837	-0.083
V3	1	15.411	0.000	1.583
V4	1	0.384	0.536	-0.250
V5	1	1.067	0.303	-0.417
V6	1	24.590	0.000	-2.000

Table 12. Simple main effects analysis (Q8)

Q8: The vibration was distracting. The results of simple main effect analysis of the independent variables in respect of Q8 are shown in Table 12. These show that there were statistically significant differences in scores between experimental and control groups in respect, V1 ($F(1,132) = 8.367, p < .05$), V3 ($F(1,132) = 15.411, p < .05$) and V6 ($F(1,132) = 24.590, p < .05$) as regards haptic effects being perceived as distracting, with experimental group users being more tolerant in this respect for V1 and V3, and the reverse holding for V6. Lastly, in the case of V2, V4, and V5, no statistically significant differences were found between the groups.

Q9: I enjoyed watching the video clip. The undertaken ANOVA highlighted that the interaction between the two independent variables (Group*Video) did not have any statistically significant effect on Q9 (Table 4). In terms of main effects, only one independent variable (video type) had a statistically significant impact on participants' enjoyment of watching video clips ($F(5,132) = 2.844, p = 0.018$). Thus, Post Hoc comparisons on Q9 using the Tukey HSD test are shown in Table 13. From the results, we can see that there is a statistically significant difference ($p = .045$) between V2 and V4 with positive (0.79) mean difference. The test also showed that V3 and V4 differed significantly ($p = .012$) with positive (0.92) mean difference. However, V1, V5, and V6 were not significantly different from the other three videos. This implies that, although there is no difference in QoE between the groups, the participants less enjoyed viewing the bright sun shining video (V4) than the blue waves (V2) and the dark solar eclipse (V3) videos.

In addition, certain association between the responses corresponding to the questions is also obtained. For example, the results above show similar responses both in the relevance of the haptic vest effects (Q2) and its impact on enjoyment

(I) Video	(J) Video	Mean Difference (I-J)	Sig.	95% Confidence Interval	
				Lower Bound	Upper Bound
1	2	-0.04	1.000	-0.82	0.74
	3	-0.17	0.990	-0.95	0.61
	4	0.75	0.068	-0.03	1.53
	5	0.13	0.997	-0.66	0.91
	6	0.14	0.996	-0.65	0.93
2	1	0.04	1.000	-0.74	0.82
	3	-0.13	0.997	-0.91	0.66
	4	0.79	0.045	0.01	1.57
	5	0.17	0.990	-0.61	0.95
	6	0.18	0.986	-0.61	0.97
3	1	0.17	0.990	-0.61	0.95
	2	0.13	0.997	-0.66	0.99
	4	0.92	0.012	0.14	1.70
	5	0.29	0.889	-0.49	1.07
	6	0.30	0.875	-0.49	1.09
4	1	-0.75	0.068	-1.53	0.03
	2	-0.79	0.045	-1.57	-0.01
	3	-0.92	0.012	-1.70	-0.14
	5	-0.63	0.196	-1.41	-0.16
	6	-0.61	0.226	-1.40	0.18
5	1	-0.13	0.997	-0.91	0.66
	2	-0.17	0.990	-0.95	0.61
	3	-0.29	0.889	-1.07	0.49
	4	0.63	0.196	-0.16	1.41
	6	0.01	1.000	-0.78	0.80
6	1	-0.14	0.996	-0.93	0.65
	2	-0.18	0.998	-0.97	0.61
	3	-0.30	0.875	-1.09	0.49
	4	0.61	0.226	-0.18	1.40
	5	0.01	1.000	-0.80	-0.78

Table 13. Post Hoc Tukey HSD test (Q9)

(Q1) in the case of V3 indicating that there is positive association between the questions for the dark-sky solar eclipse video. Similarly, in the case of V6 (bouncing balls), less enjoyment (Q1) and sense of reality (Q4) is observed in the experimental group while wearing the haptic vest. This implies that further study is required to identify the inter-relationships among the various QoE questions and with respect to the various video samples.

5 CONCLUDING DISCUSSION

Mulsemmedia is one among many technological innovations contributing to digital media. Adding more media types into mulsemmedia has the potential to enhance user QoE. However, it is important to identify a set of guidelines in creating content that leads to an actual enhancement of the user experience.

In this paper we investigate whether using crossmodal principles in creating content that addresses various senses has as effect an increase in users' QoE. Having as starting point the theory of audio-visual crossmodal correspondences

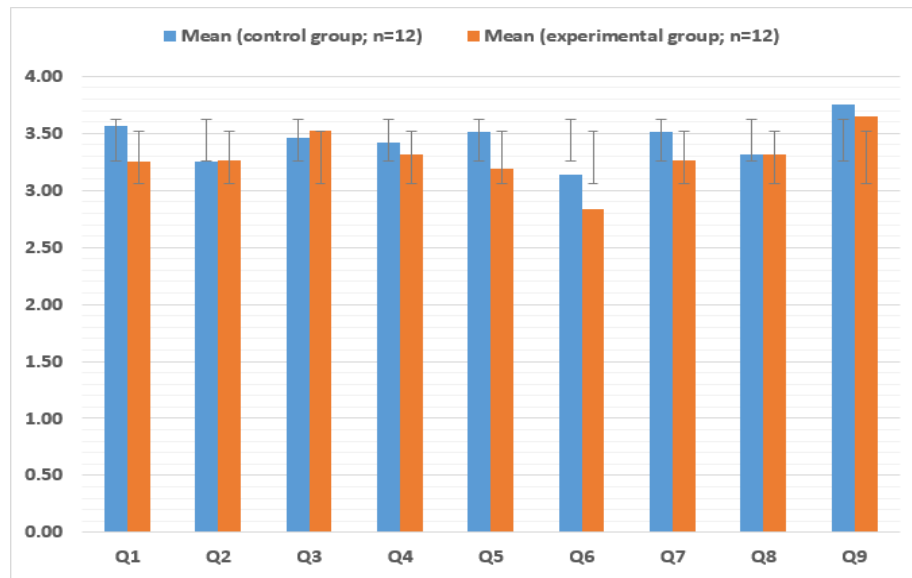


Fig. 7. Mean scores of the groups for each question

in psychology, we built audio content to match certain predominant features of the video content, and we compared the QoE with the one experienced while listening to a relaxing sound accompanying the same video.

Based on previous studies, such as [Engelke et al. 2017], we attempted to assess QoE during mulsemmedia exposure through psychophysiology-based (heart rate and eye gaze patterns) and subjective methods (Likert scale questions). In the literature, gaze tracking has been extensively investigated in the context of visual QoE assessment, the rationale being that eye gaze is directed not only by the natural content but also by potential induced distortions [Engelke et al. 2017]. In our study, we observe a totally different pattern for gazing on the screen for the individuals in the control and experimental groups (Figure 4). It is well known that eye movements depend on the visual content and on the task, but in our case the differences come from the experienced audio (and associated haptic effect). This is in line with observations from [Song et al. 2012] who showed that sound affects human gaze differently depending on the sound kind. However, the sound we use in our experiments is designed following different principles - based on crossmodal correspondences. We need to consider also that eye-tracking methodology provides in fact information about the patterns of user attention [Dumais et al. 2010]. Thus, it becomes obvious that the pattern of user attention varies depending on the type of the experienced content - crossmodally matching or not. Whilst the body of research seems to be divided on the relationship between gaze patterns and concurrent memory tasks [Gulliver and Ghinea 2004; Wang and Apperly 2017], not the same can be said in respect of the evidence that disruptions in the latter can be highlighted through different viewing patterns. Our work certainly appears to confirm this in the case of crossmodally-matched mulsemmedia.

The different viewing patterns observed in the study are corroborated to some extent by our analysis of *heart rate data*, which revealed that, with the exception of V4, the experienced high/low pitched audio (and associated haptic effect) induced a significantly higher heart rate in experimental group participants compared to those from the control group. V4 was characterized by low brightness levels, the video depicting a solar eclipse. The reason for a lower heart

rate in case of this video can be attributed to the fact that human beings are very sensitive to light exposure, and changes of light intensity can influence physiological parameters like alertness, body temperature, or heart rate [Litscher et al. 2013].

Whilst it is true that the vibrotactile effect employed in our study may affect participants' heart rate, nonetheless the effect is uniformly applied to both the experimental and control groups and any differences in heart rates between the two groups occur as a result of the different cross-modal setups between the two and not as a result of different vibro-tactile effects being applied. Accordingly, our heart rate results can be interpreted using a general rule of thumb proposed in [Gilleade et al. 2005]. In their paper, the authors analyzed videogames and proposed a game design based on high level heuristics. In their investigation, they took increases in the player's heartbeat rate as a positive sign that the player was engaging with the game. Decreases in the player's heartbeat rate were considered as a negative sign, showing lack of engagement. However, there are also other interpretations for the value of the heart rate. Sometimes this is positively associated with anxiety in general and specifically with in-game tension and negative affect [Drachen et al. 2010].

Thus, higher heart rates - associated with increased excitement, but also stress levels - in the experimental group also resulted in significantly different gaze patterns of experimental group users when compared to those associated with control group participants. Analysis of *participants' self-reported QoE* revealed that low pitched audio combined with dark featured video (in V3) induced significantly higher haptic-related QoE. However crossmodal correspondence in the case of the other videos was not demonstrated in our experiment and further research needs to elucidate crossmodal mappings in a digital context. Nonetheless, the significant influence of video clip type on user overall enjoyment (Q9) in the case of cross-modally mapped mulsemmedia is in line with other studies (which, however, did not follow cross-modal principles) [Ghinea and Ademoye 2012a; Yuan et al. 2014].

Fig. 7 shows that the average level of agreement of the experimental group of participants to self-reported questions (for most videos) is generally lower than that of the control group. This implies that respondents have enjoyed viewing the videos with original audio more than viewing them with the (crossmodally matched) high/low pitched audio. This again shows that more research needs to be undertaken to better understand the use of crossmodal correspondences in mulsemmedia. Certainly, the higher heart rates and different eye gaze patterns of users who experienced crossmodally mapped mulsemmedia alludes to increased levels of stress and excitement, which have to be investigated further. Moreover, our work also opens up opportunities for further investigation into how the incorporation of additional media types (e.g., olfaction) into mulsemmedia can be used with crossmodally matched audio-video content and, indeed, to further explore the QoE impact of even more diverse cross-modal combinations. All represent worthwhile questions which will form the focus of our future efforts.

REFERENCES

- Oluwakemi A Ademoye and Gheorghita Ghinea. 2013. Information recall task impact in olfaction-enhanced multimedia. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 9, 3 (2013), 17.
- Oluwakemi A Ademoye, Niall Murray, Gabriel-Miro Muntean, and Gheorghita 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.
- Belma R Brkic, Alan Chalmers, Kevin Boulanger, Sumanta Pattanaik, and James 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.
- Kjell Brunnström, Sergio Ariel Beker, Katrien De Moor, Ann Dooms, Sebastian Egger, Marie-Neige Garcia, Tobias Hossfeld, Satu Jumisko-Pyykkö, Christian Keimel, Mohamed-Chaker Larabi, et al. 2013. Qualinet white paper on definitions of quality of experience. (2013).
- Drew Cingel and Anne Marie Piper. 2017. How Parents Engage Children in Tablet-Based Reading Experiences: An Exploration of Haptic Feedback. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 505–510.

- 989 Anne-Sylvie Crisinel and Charles Spence. 2009. Implicit association between basic tastes and pitch. *Neuroscience letters* 464, 1 (2009), 39–42.
- 990 Edwin Dalmaijer. 2014. Is the low-cost EyeTribe eye tracker any good for research? *PeerJ PrePrints* (2014).
- 991 Fabien Danieau, Anatole Lécuyer, Philippe Guillotel, Julien Fleureau, Nicolas Mollet, and Marc Christie. 2013. Enhancing audiovisual experience with
992 haptic feedback: a survey on HAV. *IEEE transactions on haptics* 6, 2 (2013), 193–205.
- 993 Anders Drachen, Lennart E Nacke, Georgios Yannakakis, and Anja Lee Pedersen. 2010. Correlation between heart rate, electrodermal activity and player
994 experience in first-person shooter games. In *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games*. ACM, 49–54.
- 995 Susan T. Dumais, Georg Buscher, and Edward Cutrell. 2010. Individual Differences in Gaze Patterns for Web Search. In *Proceedings of the Third Symposium
996 on Information Interaction in Context (IIIX '10)*. ACM, New York, NY, USA, 185–194. <https://doi.org/10.1145/1840784.1840812>
- 997 Darragh Egan, Sean Brennan, John Barrett, Yuansong Qiao, Christian Timmerer, and Niall Murray. 2016. An evaluation of Heart Rate and ElectroDermal
998 Activity as an objective QoE evaluation method for immersive virtual reality environments. In *Quality of Multimedia Experience (QoMEX), 2016 Eighth
999 International Conference on*. IEEE, 1–6.
- 1000 Mohamad Eid, Jongeun Cha, and Abdulmoteleb El Saddik. 2008. HugMe: A haptic videoconferencing system for interpersonal communication. In *Virtual
1001 Environments, Human-Computer Interfaces and Measurement Systems, 2008. VECIMS 2008. IEEE Conference on*. IEEE, 5–9.
- 1002 Ulrich Engelke, Daniel P Darcy, Grant H Mulliken, Sebastian Bosse, Maria G Martini, Sebastian Arndt, Jan-Niklas Antons, Kit Yan Chan, Naeem Ramzan,
1003 and Kjell Brunnström. 2017. Psychophysiology-based QoE assessment: A survey. *IEEE Journal of Selected Topics in Signal Processing* 11, 1 (2017), 6–21.
- 1004 David Gal, S Christian Wheeler, and Baba Shiv. 2007. Cross-modal influences on gustatory perception. (2007).
- 1005 Gheorghita Ghinea and Oluwakemi Ademoye. 2012a. The sweet smell of success: Enhancing multimedia applications with olfaction. *ACM Transactions
1006 on Multimedia Computing, Communications, and Applications (TOMM)* 8, 1 (2012), 2.
- 1007 Gheorghita Ghinea and Oluwakemi Ademoye. 2012b. User perception of media content association in olfaction-enhanced multimedia. *ACM Transactions
1008 on Multimedia Computing, Communications, and Applications (TOMM)* 8, 4 (2012), 52.
- 1009 Gheorghita Ghinea, Christian Timmerer, Weisi Lin, and Stephen R. Gulliver. 2014. Mulsemedia: State of the Art, Perspectives, and Challenges. *ACM Trans.
1010 Multimedia Comput. Commun. Appl.* 11, 1s, Article 17 (Oct. 2014), 23 pages. <https://doi.org/10.1145/2617994>
- 1011 Kiel Gilleade, Alan Dix, and Jen Allanson. 2005. Affective videogames and modes of affective gaming: assist me, challenge me, emote me. *DiGRA 2005:
1012 Changing Views—Worlds in Play*. (2005).
- 1013 Stephen R Gulliver and George Ghinea. 2004. Stars in their eyes: What eye-tracking reveals about multimedia perceptual quality. *IEEE Transactions on
1014 Systems, Man, and Cybernetics-Part A: Systems and Humans* 34, 4 (2004), 472–482.
- 1015 Henrik Hagtvedt and S Adam Brasel. 2016. Cross-Modal Communication: Sound Frequency Influences Consumer Responses to Color Lightness. *Journal
1016 of Marketing Research* 53, 4 (2016), 551–562.
- 1017 E Hoggan and SA Brewster. 2006. Mobile crossmodal auditory and tactile displays. In *Proceedings of HAID 2006: First International Workshop on Haptic and
1018 Audio Interaction Design*. 9–12.
- 1019 Gijs Huisman, Merijn Bruijnes, and Dirk KJ Heylen. 2016. A moving feast: effects of color, shape and animation on taste associations and taste perceptions.
1020 In *Proceedings of the 13th International Conference on Advances in Computer Entertainment Technology*. ACM, 13.
- 1021 Vedad Hulusic, Gabriela Czanner, Kurt Debattista, Elena Sikudova, Piotr Dubla, and Alan Chalmers. 2009. Investigation of the beat rate effect on frame
1022 rate for animated content. In *Proceedings of the 25th Spring Conference on Computer Graphics*. ACM, 151–159.
- 1023 Vedad Hulusic, Kurt Debattista, Vibhor Aggarwal, and Alan Chalmers. 2010. Exploiting audio-visual cross-modal interaction to reduce computational
1024 requirements in interactive environments. In *Games and Virtual Worlds for Serious Applications (VS-GAMES), 2010 Second International Conference on*.
1025 IEEE, 126–132.
- 1026 Conor Keighrey, Ronan Flynn, Siobhan Murray, and Niall Murray. 2017. A QoE evaluation of immersive augmented and virtual reality speech & language
1027 assessment applications. In *Quality of Multimedia Experience (QoMEX), 2017 Ninth International Conference on*. IEEE, 1–6.
- 1028 Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis
1029 Patras. 2012. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing* 3, 1 (2012), 18–31.
- 1030 Naoya Koizumi, Hidekazu Tanaka, Yuji Uema, and Masahiko Inami. 2011. Chewing jockey: augmented food texture by using sound based on the
1031 cross-modal effect. In *Proceedings of the 8th International Conference on Advances in Computer Entertainment Technology*. ACM, 21.
- 1032 Daniela Litscher, Lu Wang, Ingrid Gaischek, and Gerhard Litscher. 2013. The influence of new colored light stimulation methods on heart rate variability,
1033 temperature, and well-being: results of a pilot study in humans. *Evidence-Based Complementary and Alternative Medicine* 2013 (2013).
- 1034 Lawrence E Marks. 1974. On associations of light and sound: The mediation of brightness, pitch, and loudness. *The American journal of psychology* (1974),
1035 173–188.
- 1036 Lawrence E Marks. 1987. On cross-modal similarity: Auditory–visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human
1037 Perception and Performance* 13, 3 (1987), 384.
- 1038 Georgia Mastoropoulou. 2007. *The effect of audio on the visual perception of high-fidelity animated 3d computer graphics*. Ph.D. Dissertation. University of
1039 Bristol.
- 1040 Georgia Mastoropoulou, Kurt Debattista, Alan Chalmers, and Tom Troscianko. 2005. Auditory bias of visual attention for perceptually-guided selective
1041 rendering of animations. In *Proceedings of the 3rd international conference on Computer graphics and interactive techniques in Australasia and South East
1042 Asia*. ACM, 363–369.
- 1043 Gene Munster, Travis Jakel, Doug Clinton, and Erinn Murphy. 2015. Next mega tech theme is virtual reality. *gene* 612 (2015), 303–6452.
- 1044 Manuscript submitted to ACM

- 1041 Niall Murray, Oluwakemi A Ademoye, Gheorghita Ghinea, and Gabriel-Miro Muntean. 2017. A Tutorial for Olfaction-based Multisensorial Media
1042 Application Design and Evaluation. *ACM Computing Surveys (CSUR)* 50, 5 (2017), 67.
- 1043 Belma Ramic-Brkic, Alan Chalmers, Aida Sadzak, Kurt Debattista, and Saida Sultanic. 2013. Exploring multiple modalities for selective rendering of
1044 virtual environments. In *Proceedings of the 29th Spring Conference on Computer Graphics*. ACM, 91–98.
- 1045 Nimesha Ranasinghe, Pravar Jain, Nguyen Thi Ngoc Tram, Koon Chuan Raymond Koh, David Tolley, Shienny Karwita, Lin Lien-Ya, Yan Liangkun, Kala
1046 Shamaiah, Chow Eason Wai Tung, et al. 2018. Season Traveller: Multisensory Narration for Enhancing the Virtual Reality Experience. In *Proceedings*
1047 *of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 577.
- 1048 Nimesha Ranasinghe, Kuan-Yi Lee, and Ellen Yi-Luen Do. 2014. FunRasa: an interactive drinking platform. In *Proceedings of the 8th International Conference*
1049 *on Tangible, Embedded and Embodied Interaction*. ACM, 133–136.
- 1050 Nimesha Ranasinghe, Thi Ngoc Tram Nguyen, Yan Liangkun, Lien-Ya Lin, David Tolley, and Ellen Yi-Luen Do. 2017. Vocktail: A Virtual Cocktail for
1051 Pairing Digital Taste, Smell, and Color Sensations. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 1139–1147.
- 1052 Han-Seok Seo, Artin Arshamian, Kerstin Schemmer, Ingeborg Scheer, Thorsten Sander, Guido Ritter, and Thomas Hummel. 2010. Cross-modal integration
1053 between odors and abstract symbols. *Neuroscience letters* 478, 3 (2010), 175–178.
- 1054 Ladan Shams, Yukiyasu Kamitani, and Shinsuke Shimojo. 2000. Illusions: What you see is what you hear. *Nature* 408, 6814 (2000), 788.
- 1055 J Simmer and V Ludwig. 2009. What colour does that feel? Cross-modal correspondences from touch to colour. In *Third International Conference of*
1056 *Synaesthesia and Art, Granada, Spain, April*.
- 1057 Ray H Simpson, Marian Quinn, and David P Ausubel. 1956. Synesthesia in children: Association of colors with pure tone frequencies. *The Journal of*
1058 *genetic psychology* 89, 1 (1956), 95–103.
- 1059 BG Slocombe, DA Carmichael, and J Simmer. 2016. Cross-modal tactile–taste interactions in food evaluations. *Neuropsychologia* 88 (2016), 58–64.
- 1060 Guanghan Song, Denis Pellerin, and Lionel Granjon. 2012. How different kinds of sound in videos can influence gaze. In *13th International Workshop on*
1061 *Image Analysis for Multimedia Interactive Services (WIAMIS 2012)*. 4–p.
- 1062 Charles Spence. 2010. The color of wine—Part 1. *The World of Fine Wine* 28 (2010), 122–129.
- 1063 Charles Spence. 2011. Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics* 73, 4 (2011), 971–995.
- 1064 Peter Stephen and Susan Hornby. 1997. Simple statistics for library and information professionals. (1997).
- 1065 Alina Striner. 2018. Can Multisensory Cues in VR Help Train Pattern Recognition to Citizen Scientists? *arXiv preprint arXiv:1804.00229* (2018).
- 1066 Yevgeniya Sulema. 2016. Mulsemmedia vs. Multimedia: State of the art and future trends. In *Systems, Signals and Image Processing (IWSSIP), 2016 International*
1067 *Conference on*. IEEE, 1–5.
- 1068 Xiuwen Sun, Xiaoling Li, Lingyu Ji, Feng Han, Huifen Wang, Yang Liu, Yao Chen, Zhiyuan Lou, and Zhuoyun Li. 2018. An extended research of crossmodal
1069 correspondence between color and sound in psychology and cognitive ergonomics. *PeerJ* 6 (March 2018), e4443. <https://doi.org/10.7717/peerj.4443>
- 1070 Benjamin Tag, Takuya Goto, Kouta Minamizawa, Ryan Mannschreck, Haruna Fushimi, and Kai Kunze. 2017. atmoSphere: mindfulness over haptic-audio
1071 cross modal correspondence. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of*
1072 *the 2017 ACM International Symposium on Wearable Computers*. ACM, 289–292.
- 1073 Atau Tanaka and Adam Parkinson. 2016. Haptic wave: A cross-modal interface for visually impaired audio producers. In *Proceedings of the 2016 CHI*
1074 *Conference on Human Factors in Computing Systems*. ACM, 2150–2161.
- 1075 R Tortell, DP Luigi, A Dozois, S Bouchard, Jacquelyn Ford Morie, and D Ilan. 2007. The effects of scent and game play experience on memory of a virtual
1076 environment. *Virtual Reality* 11, 1 (2007), 61–68.
- 1077 Augoustinos Tsiros. 2017. The parallels between the study of cross-modal correspondence and the design of cross-sensory mappings. In *Proceedings of the*
1078 *conference on Electronic Visualisation and the Arts*. BCS Learning & Development Ltd., 175–182.
- 1079 Shafiq ur Rehman, Muhammad Sikandar Lal Khan, Liu Li, and Haibo Li. 2014. Vibrotactile TV for immersive experience. In *Signal and Information*
1080 *Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. IEEE, 1–4.
- 1081 Giulio Valenti and Klaas R Westertep. 2013. Optical heart rate monitoring module validation study. In *Consumer Electronics (ICCE), 2013 IEEE International*
1082 *Conference on*. IEEE, 195–196.
- 1083 J Jessica Wang and Ian A Apperly. 2017. Just one look: Direct gaze briefly disrupts visual working memory. *Psychonomic bulletin & review* 24, 2 (2017),
1084 393–399.
- 1085 Robert B Welch, Lance D DuttonHurt, and David H Warren. 1986. Contributions of audition and vision to temporal rate perception. *Perception &*
1086 *Psychophysics* 39, 4 (1986), 294–300.
- 1087 Robert B Welch and David H Warren. 1980. Immediate perceptual response to intersensory discrepancy. *Psychological bulletin* 88, 3 (1980), 638.
- 1088 Jeffrey M Yau, Jonathon B Olenczak, John F Dammann, and Sliman J Bensmaia. 2009. Temporal frequency channels are linked across audition and touch.
1089 *Current Biology* 19, 7 (2009), 561–566.
- 1090 Zhenhui Yuan, Shengyang Chen, Gheorghita Ghinea, and Gabriel-Miro Muntean. 2014. User quality of experience of mulsemmedia applications. *ACM*
1091 *Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 11, 1s (2014), 15.
- 1092 Zhenhui Yuan, Gheorghita Ghinea, and Gabriel-Miro Muntean. 2015. Beyond multimedia adaptation: Quality of experience-aware multi-sensorial media
1093 delivery. *IEEE Transactions on Multimedia* 17, 1 (2015), 104–117.
- 1094 Longhao Zou, Irina Tal, Alexandra Covaci, Eva Ibarrola, Gheorghita Ghinea, and Gabriel-Miro Muntean. 2017. Can Multisensorial Media Improve Learner
1095 Experience?. In *Proceedings of the 8th ACM on Multimedia Systems Conference*. ACM, 315–320.