Mulsemedia: State of the Art, Perspectives, and Challenges

GHEORGHITA GHINEA, Brunel University
CHRISTIAN TIMMERER, Alpen-Adria-Universität
WEISI LIN, Nanyang Technological University
STEPHEN R. GULLIVER, University of Reading

Mulsemedia—multiple sensorial media—captures a wide variety of research efforts and applications. This article presents a historic perspective on mulsemedia work and reviews current developments in the area. These take place across the traditional multimedia spectrum—from virtual reality applications to computer games—as well as efforts in the arts, gastronomy, and therapy, to mention a few. We also describe standardization efforts, via the MPEG-V standard, and identify future developments and exciting challenges the community needs to overcome.

Categories and Subject Descriptors: H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Evaluation/methodology*; H.1.2 [**Models and Principles**]: User/Machine Systems—*Human information processing*

General Terms: Algorithms, Human Factors, Experimentation, Theory

Additional Key Words and Phrases: Mulsemedia, multisensory, contour perception, flow visualization, perceptual theory, visual cortex, visualization

ACM Reference Format:
Gheorghita Ghinea, Christian Timmerer, Weisi Lin, and Stephen R. Gulliver. 2014. Mulsemedia: State of the art, perspectives, and challenges. ACM Trans. Multimedia Comput. Commun. Appl. 11, 1s, Article 17 (September 2014), 23 pages.
DOI: http://dx.doi.org/10.1145/2617994

1. INTRODUCTION

In 2004, on the 10th anniversary of the creation of the ACM Multimedia Special Interest Group, Larry Rowe and Ramesh Jain published a seminal paper "Future Directions in Multimedia Research" in ACM TOMCCAP. The paper presented the result of discussions at a one-day workshop with over 30 leading researchers in the field. There was agreement that multimedia is a multidisciplinary field, applied to a variety of fields (e.g., entertainment, education, medicine, creative arts, etc.). Three unifying themes were identified to unite the multimedia research field. First, a multimedia system or application is comprised of at least two media objects that are correlated. Second, there is the issue of integration and adaptation where multiple media objects should be used jointly and separately to improve application performance, and distributed multimedia applications should provide transparent delivery of dynamic content in such a way that

Authors' addresses: G. Ghinea (corresponding author), Department of Computer Science, Kingston Lane, Uxbridge, UB8 3PH, U.K.; email: george.ghinea@brunel.ac.uk; C. Timmerer, Universitätsstrasse 65-67 A-9020 Klagenfurt Austria; email: christian.timmerer@itec.uni-klu.ac.at; W. Lin, School of Computer Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798; email: wslin@ntu.edu.sg; S. R. Gulliver, Henley Business School, Whiteknights, Reading, RG6 6UR, U.K.; email: s.r.gulliver@henley.reading.ac.uk.

DOI: http://dx.doi.org/10.1145/2617994

content adapts naturally to the users' environment. Third, multimedia applications are multimodal and interactive [Rowe and Jain 2005].

Ten years later what has changed? A lot, and maybe not so. Arguably, the three unifying themes are very much valid today, in a world dominated by social media and a proliferation of sensor-rich (predominantly mobile) devices, where individuals are producers, broadcasters, and consumers of rich media content. Reassuringly, the accepted definition of multimedia remains that of a combination of two or more media, one of which is preferably continuous, the other usually discrete. It is without doubt that most of multimedia content available today are a combination of video and audio (both continuous media) with textual (discrete media) information sometimes contained therein. However, such applications engage primarily two of our human senses: that of sight and hearing, that is, they are bisensorial. This situation is at odds with the fact that 60% of human communication is nonverbal and that most of us perceive the world through a combination of five senses (i.e., sight, hearing, touch, taste, and smell). As such, current multimedia experiences fail to convey the sensation, for instance, of heat and humidity, let alone the wafts of aromas that one experiences when waking through a spice market in India. As humans, we engage and learn by interacting with all of our senses—can we not do this in a digital fashion as well?

We therefore propose mulsemedia—multiple sensorial media—as a new multimedia challenge for the forthcoming ten years. Whereas multimedia applications are usually bi-(sometimes tri-)media and almost exclusively bisensorial in nature, mulsemedia applications are those that engage three (or more) of our senses.

While current technological developments have made *digital* mulsemedia experiences somewhat of a novelty, in the nondigital world, they are anything but. The earliest we know of happened in 1906 when artificially-generated smells were combined with audiovisual content. An audience was sprayed with the scent of roses while watching a screening of the Rose Bowl football game. In 1943, Hans Laube who had earlier perfected a technique for extracting odors from an enclosed environment, was able to reverse this process so that selected scents were emitted at specific times and for specified durations, resulting in a 35 minute 'smell-o-drama' movie called *Mein Traum* in which 35 different odors were released to accompany the drama presentation. Building on this, audiences in 1959 viewing a documentary about Red China called *Behind the Great Wall* were treated with an AromaRama presentation, in which the theatre's air-conditioning system was used to release over 30 different smells. Shortly afterwards, in 1960, Michael Todd Jr., produced a competing system called *Smell-O-Vision,* in which aromas were released during the screening of the movie *Scent of Mystery*. It would be an exaggeration to say that these experiences were an unqualified success: challenges of generating realistic scents, the tendency of odors to drift and diffuse, as well as insufficiently understood characteristics of odor intensity all meant that, novelty factor aside, user take-up was low. The reaction of the audience to the AromaRama experience is probably best described from the following extract from the review published back then by *Time* magazine.

> "To begin with, most of the production's 31 odors will probably seem phoney, even to the average uneducated nose. A beautiful old pine grove in Peking, for instance, smells rather like a subway rest room on disinfectant day. Besides, the odors are strong enough to give a bloodhound a headache. What is more, the smells are not always removed as rapidly as the scene requires: at one point, the audience distinctly smells grass in the middle of the Gobi desert."

Such drawbacks did not prevent pioneering mulsemedia efforts, however. In 1962, Morton Heilig created what is now popularly dubbed as the first virtual reality (VR) experience for users, even though digital computing and virtual reality systems did not exist then. With *Sensorama*, he created an arcade-style device, which took users on an immersive 3D virtual reality bike ride experience through the streets of Brooklyn, New York. This came complete with motions and vibrations, sounds, fans, and smells, the most complex mulsemedia experience devised so far, engaging four out of our five major senses. Indeed, given that the sense of taste is intimately connected to that of smell and that one of the aromas emitted was that of freshly baked bread from a bakery, it is not inconceivable that for some users all five major senses were engaged in their mulsemedia journey [Heilig 1962].

Over half a century has passed since then, so where are we now on the mulsemedia landscape? To answer the question, this article reviews developments that recent technological advances have made possible in order to see how mulsemedia applications fit within the multimedia arena, and to identify challenges that the community has yet to overcome. Accordingly, the structure of the rest of this article is as follows: given the importance of the human sense to mulsemedia, the next section gives an overview; Section 3 then details related work. Mulsemedia needs standards to thrive, and to this end, Section 4 describes MPEG-V, a standard capable of supporting mulsemedia applications. The user is an important element of mulsemedia, and QoE efforts in this respect are detailed in Section 5. Finally, research challenges and open issues are described in Section 6.

2. HUMAN SENSORIAL OVERVIEW

In this section, we consider in more detail the multiple process steps required to achieve multiple sensory perception. We introduce key physiological systems and describe how each captures and transforms information from the world so that the brain can process it. We conclude the section by considering the issue of cognitive binding and highlight the attentive struggle between top-down and bottom-up processes.

2.1. Multiple Sensory Perception

Sensory perception relates to a human's conscious sensory experience of the world, that is, what a person can see, hear, smell, touch, and taste, etc. When we consider mulsemedia perception, therefore, it is critical to appreciate that multiple sensory media perception is not something that just 'happens.' For a person to be able to understand and assimilate meaning from multiple sensory media, they must capture, interpret, and combine information from numerous sensory organs, that is, bottom-up sensing [Goldstein 2013]. Moreover, information from multiple senses must be cognitively joined and aligned and then compared to higher-order cognitive schema, which define task semantics, pragmatics, and social norms, that is, top-down thinking [Marois and Ivanoff 2005; Mayer 2003].

Although perception sometime feels as though it just happens, it is, in reality, the result of a complex set of processes. Biological sensors capture physical signals from the environment and transduce them, with the exception of specific chemoreceptors, into structured electrical signals. These signals are restructured in the nervous systems and transmitted to the brain. Within the brain, spatial/temporal signals are then subconsciously structured as patterns, which are attentively processed as higher-level artefacts/objects. Once structured, appreciation of meaning facilitates the validation of propositions. Identifying whether something is true or false facilitates humans to align bottom-up sensory input with top-down knowledge and memory, and enables us to create and iteratively validate complex schema models of the real world. The existence of these complex schema models allows humans to predict and understand the world in the context of higher pragmatic and social structures.

2.2. Capturing the Physical

There is no universal agreement as to the number of senses perceived by the human mind. In reality, however, the human body manages sensory inputs from a wide range of internal and external sensory inputs, such as pain (nociception), space (proprioception), movement (kinaesthesia), time, and temperature (thermoception). As well as external senses, our bodies sense and process internal regulation (called interoceptive senses), which leads to feelings of hunger, sickness, thirst, stress or discomfort [Craig 2003]. All of these senses are internally linked within our model of the world. Despite our processing this dynamic range of internal and external senses, mulsemedia systems focus on the five traditional sense, as defined by Aristotle, that is, visual (sight), auditory (sound), tactile/haptic (touch), olfactory (smell), and gustatory (taste). In this section, we introduce the reader to each physiological system in turn.

2.2.1. Visual (Sight). In mulsemedia, sight allows assimilation of textual and visual information. Light reflected from a physical object in the visual field enters the eye through the pupil and passes through the lens, which projects an inverted image onto the retina at the back of the eye. The retina consists of approximately 127 million light-sensitive cells (120 million called rods; 7 million called cones, which can be subdivided into L-cones, M-cones, and S-cones). Although cones are less light sensitive than rods, they are responsible for capturing color within the human visual system. When light enters the eye, it passes through seven sensory cell-layers before reaching the rods and cones at the back of the eye. If cones were distributed evenly across the retina, their average distance apart would be relatively large, leading to poor spatial acuity. Accordingly, cones are concentrated in the center of the retina (in a circular area called the macula lutea). Within this area, there is a depression called the fovea, which consists almost entirely of cones, and it is through this area of high acuity, extending over just 2° of the visual field, that humans make their detailed observations of the world. The cells that process and transmit information to the brain are called the bipolar, horizontal, and ganglion cells. Photoreceptors at the back of the eye (cones and rods) are activated when light is shined at them, which consecutively activates bipolar cells. Visual pre-attentive segregation and object combining occur primarily in the occipital lobe (at the back of the brain); however, visual information is contextualized, that is, 'where/how' and 'what,' in the parietal and temporal lobes, respectively [Schiller 1986].

2.2.2. Auditory (Sound). In mulsemedia, the human auditory system is used heavily in the transfer of sound, speech, music, and special effects. If an object vibrates, it produces a sequence of wave compressions in the air surrounding it. These fluctuations in air pressure spread away from the source of vibration at 320m/s, reducing in magnitude as the energy is dispersed. When two or more waveforms interact, they create a combined waveform that is the sum of its component parts. Sound is the sensation produced by the ear when a vibration occurs within a given frequency range (approximately 20Hz to 20KHz), which is audible to humans. The volume of sound, at the source of vibration, is dependent upon the magnitude of the sound energy waveform. The frequency is dependent upon the frequency of compressions being produced by the source of vibration.

The ear is divided into three parts—the outer (external), the middle, and the inner (internal) ear. The outer ear collects sound waves and focuses them along the ear canal to the eardrum. The eardrum vibrates, causing bones (malleus and incus) to rock back and forth, which passes movement to the cochlea where fluid in the inner ear is disturbed. The disturbance of fluid causes thousands of small hair cells to vibrate. The cochlea converts sound waves into electrical impulses, which are passed on to the brain via the auditory nerve. The three main auditory areas in the brain (i.e., the core area, the belt area, and the parabelt) are found in the temporal lobe. Recognition of sound and localization of sound are, however, processed separately [Yost and Nielsen 1985].

2.2.3. Tactile/Haptic (Touch). In mulsemedia, tactile feedback allows us to identify several distinct types of sensations; as human skin contains a number of different sensory receptor cells that respond preferentially to various mechanical, thermal, or chemical stimuli. The majority of multimedia studies involves the tactile or touch sense, which detects pressure and touch (i.e., brushing, vibration, flutter, and indentation); however, human skin is also sensitive to temperature and pain. Information from the skin receptors is carried along a “touch-neuron pathway” to the somatosensory cortex, which maps the senses in the body and transmits messages about sensory information to other parts of the brain (e.g., for use in performing actions, for making decisions, enjoying sensation, or reflecting on them).

2.2.4. Olfactory (Smell). In mulsemedia, olfactory feedback allows researchers to monitor subconscious reaction to smell, which is often linked to task/emotional contextualization. There are 50 million primary sensory receptor cells in a small (2.5cm^2) area of the nasal passage called the olfactory region. The olfactory region is formed of cilia projecting down out of the olfactory epithelium into a layer of mucous, which helps to transfer soluble odorant molecules to the receptor neurons. The neuronal cells form axons, which penetrate the cribriform plate of bone, thus reaching the olfactory bulb of the brain. Smell messages are sent directly to the higher levels of the central nervous system via the olfactory tract, where olfactory information is decoded and a reaction is determined. Compared to many mammals, smell ability in humans is limited. Smell is, however, important to human perception of episodic knowledge, with smells often triggering specific contextual memories. The olfactory sense is used in humans as a means of identifying resources, as a warning of danger (e.g., rotten food, chemical dangers, and fires), identify mates, predators, aiding navigation, and providing sensual pleasure. Since olfactory neurons are connected directly to the brain and can therefore unconsciously influence cognition and emotion, smell is known to trigger discomfort, sympathy, or even unconscious refusal [Ayabe–Kanamura et al. 1998].

2.2.5. Gastronomy (Taste). The tongue is covered in papillae, which are either (i) filiform, found across the entire surface of the tongue, (ii) fungiform, which are found on the tip and sides of the tongue, (iii) foliare, which are structured at the sides at the back of the tongue, or (iv) circumvilliate, found at the central back of the tongue. All papillae, with the exception of filiform, contain taste buds. Each of the 10,000 taste buds contains between 50–100 taste cells. Traditionally it was believed that taste was grouped in areas relating to sour, sweet, salty, bitter (with umami not considered); however, it is now understood that all tastes (including umami) are registered by all taste buds. Electrical signals are generated in taste buds and pass along one of a number of nerves, relating to separate areas of the tongue, and link to both the thalamus (perched on top of the brainstem) and the frontal lobe.

Smell and taste are commonly considered together, as they are functionally linked. Unlike other senses, which interpret light/sound waveforms or interaction patterns and transform these into electrical signals understood in the brain, smell and taste are often termed ‘gatekeeping’ senses, that is, sensations created as a result of interaction with molecules being assimilated into the body [Goldstein 2013]. Gatekeeper (chemoreceptor) senses are understandably linked to biological and emotional processes, that is, to ensure automatic rejection in the case of bad food. Despite input of data via separate sensory systems (i.e., smell and taste), it is almost impossible to taste something while pinching your nose, making the experiences of smell and taste hard to separate.

2.3. Binding and Focus

Although entities and events in the world are perceived via disparate sensory modalities, as described in Section 2.2, our experience of the world is largely coherent (both spatially and temporally). The issue of how the brain integrates and aligns sensory fragments is called binding [Damasio 1989, p. 29], and consists of segregation and

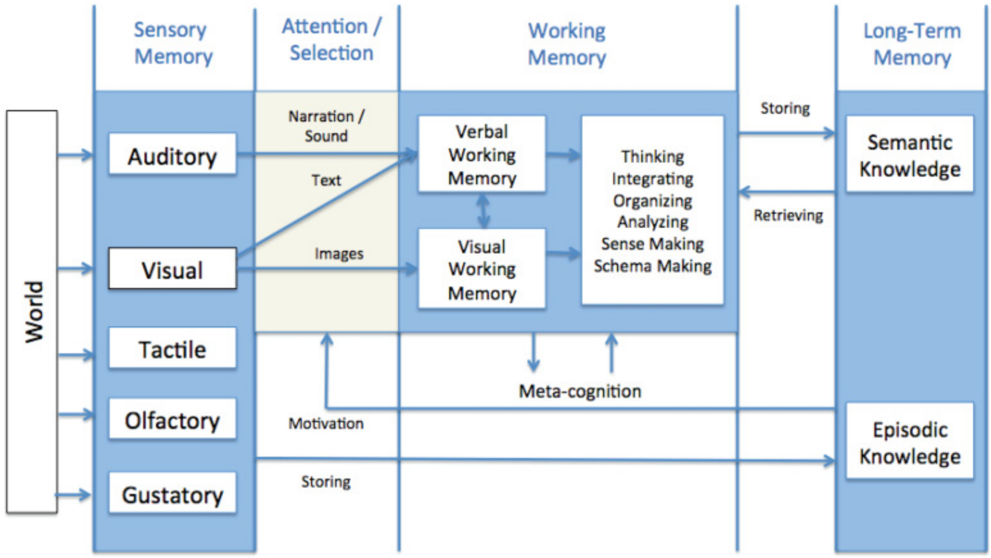

Fig. 1. Schema of thinking processes (based on [Mayer 2003; Marois and Ivanoff 2005]; adapted from [Fadel and Lemke 2008].

combining processes. Segregation processes (BP1) define high-level object variables within each sensory input (e.g., shape and color from the same input from millions of light-sensitive cells), and combining processes (BP2) relates to the process of joining and synchronizing object variables across different senses. Sensory processing is consistent for all humans, and researchers understand a significant amount concerning the processing and representation of sensory data [Smythies 1994, p. 54]; however, there is less understanding of how brain mechanisms construct phenomenal objects (i.e., high-level mental object, either physical or conceptual, which acts as the focus of attention).

Dual-process theory, which provides some interesting insights, separates cognition into two systems, that is, intuition/experience (termed system 1) and reasoning/memory (termed system 2). System 1 combines sensory and emotional stimuli to subconsciously define spatial/temporal associations between object variables. System 2 allows humans to uniquely process conscious judgments and attitudes in the context of semantic and episodic knowledge. System 2 is slow, however, and limited in part due its reliance on limited-capacity serial-based memory functions (see Figure 1.). Such limitations mean that conscious perception occurs in linear installments, with task efficiency significantly reduced if multitask switching is required [Rubinstein et al. 2001]. Due to its limitations, human reasoning is significantly dependent upon existing knowledge (schema) to support simplification of the task or contextualize episodic information, and has been shown to influence user attentive selection [Yarbus 1967].

2.4. Summary

It is clear that mulsemedia sensory media perception is not something that just happens. Perception is a complex combination of steps that combine bottom-up (sensory processing) and top-down (cognitive reasoning) processes, which results in the appreciation of the media information and the interpretation of its meaning in the context of existing semantic and episodic knowledge. Sensory processing is well understood. The process of understanding how knowledge impacts mulsemedia media interpretation, perception, and acceptance, however, is an exciting area of research.

3. RELATED WORK

Mulsemedia research, while not mainstream and sheltering perhaps under more traditional research areas, has nonetheless progressed over the past 20 years. In this section, we present key work in the area. We start off by highlighting work done on mono-sensorial evaluation. Of course, most work performed so far in this respect targets audition and vision. Since the emphasis of this article is on mulsemedia, we will not discuss in detail perception-based models for speech, audio, image, graphics, and video; interested readers can refer to the recent surveys in such modeling and applications (e.g., [Campbell et al. 2009; Lin and Kuo 2011; Möller et al. 2011; Reinhard et al. 2013; Richard et al. 2013; Wu et al. 2013; You et al. 2010]). Therefore, the main thrust of Section 3.1 is to introduce existing research involving other senses, namely, olfaction, taction, and gustation, but not in combination with one another. The Section 3.2 then proceeds to review research exploring the combination of two or more senses in a digital environment: mulsemedia, while the Online Appendix gives more details for the related basic technical approaches and computational models so far in the literature, although the development of mulsemedia algorithm and systems is still in its infancy.

3.1. Mono-Sensorial Evaluation

There has been interesting and substantial research into the olfactory system enabling humans to recognize and categorize different odors and determining many behavioral and social reactions. Ho and Spence [2005] investigated the differential effects of olfactory stimulation under conditions of varying task difficulty. Participants detected visually-presented target digits from a stream of visually-presented distractor letters in a rapid serial visual presentation task; at the same time, participants were required to discriminate stimuli presented on the front or back of their torso. The results showed a significant performance improvement in the presence of peppermint odor (as compared to air) in a difficult task but not in an easy one. This demonstrated that olfactory stimulation can facilitate tactile performance.

In the digital world, a pioneer in the area of olfaction is Kaye [2001], who, in his work on symbolic olfactory devices, experimented with a few prototypical designs of olfactory data display devices to illustrate the concept of computer-controlled smell output. For human beings, odor stimuli are highly associated with many processes, such as emotions, attraction, mood, etc. Monitoring and analyzing electroencephalogram (EEG) of human brain activity during perception of odors has shown [Yazdani et al. 2012] that classification of EEG signals during perception of odors can reveal the pleasantness of the odor with relatively high accuracy. Ghinea and Ademoye [2010a, 2011] focused on olfaction-enhanced applications. The challenges of enhancing mulsemedia with olfaction were also discussed.

Taction is another important sense for mulsemedia investigation. For foundational knowledge in this area and guidelines of design, readers can refer to the paper Seungmoon and Kuchenbecker [2013]. Haptic rendering (or haptic display) conveys information about virtual objects to users through the sense of touch. For haptic rendering, force-feedback display of contact interactions can be realized for both rigid and deformable virtual models. A general framework for force-feedback display of virtual environments is presented by Otaduy et al. [2013], and the issues, modeling, and assessment related to haptic aesthetics is discussed by Carbon and Jakesch [2013]. It has been shown that perceiving material properties (including roughness, friction, and thermal properties) of objects through touch is generally superior to the perception of shape [Klatzky et al. 2013].

With respect to gustation, this sense is intricately linked with olfaction. However, the only work targeting gustation *per se* of which we are aware is that of Adrian Cheok (`http://adriancheok.info`). He and his team developed a taste transmitter machine for sending tastes remotely (the user sticks his/her tongue in a device which transforms a signal delivered over the Internet into electrical impulses to the tongue). Coupled with the group's work in developing a machine for sending olfactory signals over a network, the ultimate aim is to build a world repository of gastronomic knowledge, presumably accessible online to users everywhere.

3.2. A Review of Mulsemedia Research and Applications

Mulsemedia research is usually inextricably linked to the development of novel and exciting applications. One of the earliest such mulsemedia VR application is that of Cater [1992] and his team, who developed a virtual reality system to train potential fire-fighters to recognize characteristic smells commonly associated with fires. The problem being solved in this case was to familiarize potential fire-fighters with those smells that are often associated with fires, as it is often thought and argued that it is easier to recognize smells already known by a person. Moreover, in a fire-fighter's profession, being able to detect the presence of such smells could well prove invaluable.

Later on, Dinh et al. [1999] investigated the use of tactile, olfactory, and auditory sensory modalities with different levels of visual information on a user's sense of presence and memory of the details of a virtual reality experience. With respect to the olfactory sensory modality, the research study was limited when compared with other sensory modalities considered. Moreover, the single olfactory cue used in the study did not produce any significant effect on the sense of presence, although it did on memory.

One of the benefits of integrating mulsemedia interfaces in applications is that it can overcome literacy barriers and bring the world of computing closer to categories of people who had hitherto been excluded from it. Jain [2003] was one of the earliest to make this point when describing the potential of Experiential Computing—computing based on the way humans naturally experience and interact with their environment. Based primarily on video, audio, and taction, he then describes the potential that such interfaces might have in enhancing virtual and augmented reality systems.

In related work, Bodnar et al. [2004] created a notification system that made use of mulsemedia data. In their work, they conducted an experimental study to compare the effect of visual, audio, or olfactory displays the delivery notifications had on a user's engagement of a cognitive task. Participants were given an arithmetic task to complete, and at various intervals, two types of notifications were triggered: (1) participants had to immediately stop what they were doing and record some data before returning to the completion of their task, and (2) they could ignore the notification. With this experiment, they found that while olfactory notifications were the least effective in delivering notifications to end users, they had the advantage of producing the least disruptive effect on a user's engagement of a task. It is also worth noting that they encountered most of the problems of using smell output as highlighted earlier by Kaye [2001] and had participants mostly commenting that some of the smells used were too similar to be distinguishable. Lingering smells in the air also made it difficult to detect the presence of new smells, and the lack of experience of working with olfactory data impacted their performance of the assigned task.

Brewster et al. [2006] used explicitly-learned odor memories to evaluate the effectiveness of using olfactory data to aid in multimedia content searching, browsing, and retrieval in a digital photo library. To conduct this experiment, they developed an olfactory photo browsing and search tool, which they called Olfoto. The odors are learned by getting participants to complete the explicit odor memory task of associating specific odors with their personal photographs, that is, smell-based photo tags. Participants were also required to tag the same photographs using text-based tags. The testing phase occurred two weeks later, in which participants were asked to complete three types of exercises.

Two were matching exercises that required matching photos with the smell/text tags they had previously associated with them—in one exercise, multiple photos were presented with one smell/text, tag, and in the other, multiple smell/text tags were presented with one photograph. The third exercise involved searching through their digital photo libraries using smell or text tags after being given three key features of the photo. Despite the fact that research has shown that odor memories persist longer than word and verbal memories, the results showed that performance was lower with the smell-based tags. The lower performance may well be attributed to the fact that, possibly, odor memories linked to emotions, that is, those implicitly learned last longer than those explicitly learned. Thus while in this experimental study participants learned to associate an odor memory with their photographs, the memory was probably not as profound as it would have been if the odor memory had been implicitly learned during the real-life moment when the photo was taken. Nonetheless, the findings from their study suggest that odor memories do have the potential to play a role in multimedia content searching.

In related work, the effects of olfaction on information recall in a virtual reality game environment were evaluated by Tortell et al. [2007]. In this experimental study, participants engaged in game play in a virtual reality environment. The first phase of the study involved an implicit odor-learning period for one group of participants, where subjects had a smell present whilst playing the virtual reality game. The other group of participants in this phase of the experiment had no smell present while they played the game. In the second phase of the experiment, which was an information recall task about the VR environment, participants were again split into two groups. One group performed the task with the same smell that was present during the first phase of the experiment, while the second group performed the information recall task with no smell present. Participants were randomly assigned to groups in the two phases of the experiments, so that participants who completed the first phase of the experiment in the presence of smell did not necessarily get to complete the second phase with the presence of smell and vice versa. Results showed that the subjects who were presented with scent only during the recall phase performed by far the worst, while subjects with scent only during the VR experience performed the best. However, the general findings from the study did show that the introduction of scent in the VR environment had a positive effect on subjects' recollection of the environment.

Multimedia entertainment, such as computer games, is another area that is expected to benefit from the addition of our other sensory cues (thus becoming mulsemedia games). It is expected that they will heighten the sense of presence and reality and hence impact positively on user experience, (e.g., make it a more engaging experience for users). We next mention some media entertainment systems that involve the use of olfactory data in one way or another.

Fragra is a visual-olfactory virtual reality game that enables players to explore the interactive relationship between olfaction and vision [Mochizuki et al. 2004]. The objective of the game is to identify if the visual cues experienced correspond to the olfactory cues at the same time. The game environment has a mysterious tree that bears many kinds of foods. Players can catch these food items by moving their right hand and when they catch one of the items and move it in front of their nose, they smell something which does not necessarily correspond to the food item they are holding. Although they do not report on any detailed evaluation of their implemented game, they do report that in their preliminary experiment, the percentage of questions answered correctly varied according to the combination of visual information and olfactory information and conclude that there is a possibility that some foods' appearance might have stronger information than their scents, and vice versa. A similar interactive computer game called the “Cooking Game” was created by Nakamoto and his research team at the Tokyo Institute of Technology [Nakamoto et al. 2008].

In earlier related work, Boyd-Davis et al. [2006] used olfactory data to create an interactive digital olfactory game. However, the main objective of their experiment, “What should the designer of interactive systems know about olfactory data?” is a question already answered by predecessors in the field. In their work, they developed a suite of digital games in which they used olfactory data, (i.e., three different scents) to engage users in game play. The users’ sense of smell is the main skill needed to win the games. The findings from their work further confirm results reported by Kaye [2001] about the use of olfactory data.

Morrot et al. [2001] carried out a similar study to investigate the interaction between the vision of colors and odor determination using lexical analysis of wine experts’ tasting comments. For the experiment, they simulate a wine tasting practice, where the wine tasters provide comments on the tasted wines based on the visual, olfactory, and gustatory properties of the wines. A previous study [Williams et al. 1984] had actually shown that perception of the olfactory qualities of wines changes depending on whether the color of the wine is visible or hidden from the subjects by using transparent and opaque wine glasses, respectively. In the study carried out by Morrot et al., they colored a white wine artificially red and presented it to wine experts to analyze, alongside the uncolored white wine and a red wine. To confirm that the colorant used to artificially color the wine had no influence on the colored wine, a pre-test experiment was carried out to confirm that the white wine and its artificially colored version were perceived as the same when its color was obscured from the tasters. Their results showed that the white wine was perceived as having the odor of a red wine when colored red (all of the wine tasters that participated in the study described the artificially colored wine with terms relating to red wine qualities; the wine’s color thus appears to provide significant sensory information, which misleads the subjects’ ability to judge flavor; lastly, the mistake is stronger in the presence rather than in the absence of access to the wine color).

The Research in Augmented & Virtual Environment Systems (RAVES) research group reported a study conducted to investigate the impact of olfaction (concordant and discordant scents) on a user’s sense of immersion into a virtual reality environment [Jones et al. 2004]. The experimental study involved participants playing a computer game in an immersive virtual experiment. The experimental conditions consisted of a control case where no scents were released while the participant played the game and two experimental cases, one involving concordant scents (e.g., emission of an ocean mist scent as the player passed the ocean and a musty scent when the player was in the fort in the immersive environment) and the other a discordant scent (e.g., smell of maple syrup throughout the game). The results from this study were not statistically significant, however.

It is of little surprise that because of the relative novelty of the mulsemedia combinations involved, the studies reviewed so far also explore user acceptance of these new media objects. This is a theme carried forward in more recent research [Ghinea and Ademoye 2012], which looked at user perception and acceptance of olfactory media combined with the more traditional audio and video.

Kahol et al. [2006] present strategies and algorithms to model context in haptic applications that allow users to explore haptic objects in virtual reality/augmented reality environments. The results from their study show significant improvement in accuracy and efficiency of haptic perception in augmented reality environments when compared to conventional approaches that do not model context in haptic rendering. Indeed, the use of haptics in mulsemedia VR environments has recently been the subject of other research (e.g., [Apostolopoulos et al. 2012]).

In related work, researchers reported on a perceptual study carried out to establish an algorithm to provide high-quality intermedia stream synchronization between haptic and audio (voice) media objects in a virtual environment [Ishibashi et al. 2004]. Indeed, synchronization seems to be a common theme across mulsemedia research. Thus, recent work has explored synchronization of olfactory media with audiovisual content [Ghinea and Ademoye 2010a], while Steinbach et al. [2012] investigated synchronization issues between different modalities and the integration of video and haptics in resource-constrained communication networks. Ghinea and Ademoye [2010b] tackled olfaction-enhanced mulsemedia by combining computer-generated smell with haptic data.

Interactive media and applications have become ubiquitous and compete for attention in our everyday life and work. As discussed by Sarter [2013], this ubiquity has led to an increasing need for effective multimodal interfacing and decisions, including information distribution across different sensory channels to ensure detection, interpretation, and handling of signals. An overview of well-known models of multimodal management was presented by Sarter. In related work, Gray et al. [2013] presented studies of multisensory (audio, tactile, etc.) integration and cross-modal spatial attention to engage more than just a single sense in complex environments. First, multimodal signals were used to reorient spatial attention under the conditions in which unimodal signals may be ineffective. Second, multimodal signals are less likely to be masked in noisy environments. And last, natural links exist between specific signals and particular behavioral responses. A multimodal system should be designed to minimize any incongruence presented in different sensory modalities that relate to the same event.

We also mention that mulsemedia has great therapeutic potential. While aromatherapy, music therapy, and therapies based on touch all employ primarily one human sense, the creation of multisensory rooms, which give mulsemedia experiences to individuals with special needs, ranging from learning difficulties to autism, Alzheimer's, and dementia, has been reported. Accordingly, the EU Framework Project 5 *MEDIATE* reported research on rooms comprising both visual (e.g., light, color, UV light, projections, illusions), audio (e.g., soothing music), olfactory (i.e., aromatherapy dispensers), and tactile stimuli (i.e., objects with different textures, shapes, vibration) [Gumtau 2011]. Across the Atlantic, and again for therapeutic purposes, *Multisensory Systems* (`http://multisensorysystems.com`) have developed an immersive mulsemedia system integrating 3D sound, olfaction, vibration, and imagery.

Last but not least, mulsemedia applications were first created in association with the film industry. So it should come as no surprise that the arts and creative industries continue to experiment with mulsemedia in their content and delivery mechanisms. In so doing, interactive digital experiences are no longer audiovisual creations but mulsemedia ones. The integration of haptic and olfactory capabilities in many contemporary interactive designs makes the communicative potential of mulsemedia in terms of sensory, affective, individual, and creative expression even more relevant. Thus for instance, Bamboozle theatre (`http://www.bamboozletheatre.co.ok`) and Oily Theatre (`http://www.oilycart.org.uk`) both specialize in multisensory performances tailored exclusively for children with autism or complex disabilities. Theatrical mulsemedia experiences are also for mainstream audiences—Disney's 4D movie experiences featuring tactile and olfactory stimuli on top of the traditional audiovisual presentation have been a staple of audiences for the last 30–40 years. Dynamic Motion Rides (*DyMoRides*) is an Austrian company that has developed a host of "complex and innovative entertainment attractions," all involving mulsemedia, for a wide range of entertainment parks worldwide; while the well-known Lowry theatre in Manchester will be staging Nosferatu (`http://www.thelowry.com/event/nosferatu`), a mulsemedia theatrical event in February 2014, no less.

4. MPEG-V: A STANDARD FOR MULSEMEDIA

4.1. Context and Objectives

The initial purpose of the MPEG-V standard was to provide an architecture and associated information representations to enable the interoperability between virtual worlds and the real world. This also explains the name MPEG-V, where “V” stands for virtual world and the standard was entitled “information exchange with virtual worlds,” later renamed to “media context and control” to broaden its scope.

The actual architecture of the MPEG-V standard defines interfaces, which are provided in the form of XML- and binary-based representation formats, between digital content providers (including virtual worlds) and real-world devices comprising sensors and actuators. These real-world devices may offer various capabilities controlled by appropriate device commands issued by the digital content applications. Alternatively, these commands may be also used to control devices within virtual worlds.

The MPEG-V standard comprises the following parts.

—*Part* 1. Architecture—describes the general system architecture as well as major interfaces and interoperability points.
—*Part* 2. Control Information—defines the means to describe the capabilities of (real-world) devices as well as to control them.
—*Part* 3. Sensory Information—provides the means to describe sensory effects as discussed in the next section.
—*Part* 4. Virtual World Object Characteristics—provides data representation formats to specify virtual objects that can be exchanged with other virtual worlds.
—*Part* 5. Data Formats for Interaction Devices—focuses on device interactivity and associated data formats.
—*Parts* 6 and 7. Define common data types and tools needed for the other parts as well as conformance and reference software.

4.2. Sensory Information

The main purpose of MPEG-V Part 3—Sensory Information—is to enhance both the quality and user experience of multimedia services by annotating existing multimedia content with additional sensory effects. The main motivation behind this work is that the consumption of multimedia content may stimulate also other human senses—going beyond hearing and seeing—including olfaction, mechanoreception, thermoception, etc. Therefore, multimedia content is annotated providing so-called sensory effects that steer appropriate devices capable of rendering these effects, giving the user the sensation of being part of the particular media which results in a worthwhile, informative user experience.

4.2.1. Concept and System Architecture. The concept and system architecture of receiving sensory effects in addition to audio/visual content is depicted in Figure 2. The media and the corresponding sensory effect metadata (SEM) may be obtained from a digital versatile disc (DVD), Blu-ray disc (BD), or any kind of online service (i.e., download/play or streaming). The media processing engine, which can be deployed on a set-top-box, DVD/BD player, or any other smart device, is responsible for playing the actual media resource and accompanying sensory effects in a synchronized way based on the user’s setup in terms of both media and sensory effect rendering. Therefore, the media processing engine may adapt both the media resource and the SEM according to the capabilities of the various rendering devices.

The MPEG-V Part 3 standard deliberately defines only the representation formats without detailing how to create and how to consume multimedia content enriched with sensory effect metadata. This approach enables interoperability among different

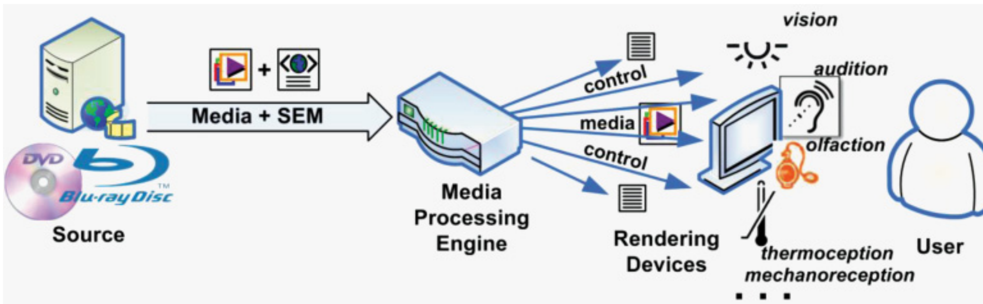

Fig. 2. Concept and system architecture of sensory information.

vendors while supporting a broad range of application domains. Possible means for creating and consuming multimedia with sensory effects, including its quality assessment, are described in Section 5.

The representation formats defined within MPEG-V Part 3 are now described in the following.

4.2.2. Sensory Effect Description Language (SEDL). The Sensory Effect Description Language (SEDL) is an XML Schema-based language which enables one to describe so-called sensory effects such as light, wind, fog, vibration, etc., that trigger human senses. The actual sensory effects are not part of SEDL but defined within the Sensory Effect Vocabulary (SEV) for extensibility and flexibility allowing each application domain to define its own sensory effects (see Section 4.2.3). A description conforming to SEDL is referred to as Sensory Effect Metadata (SEM) and may be used in any multimedia content (e.g., movies, music, websites, games). The SEM can steer sensory devices like fans, vibration chairs, lamps, etc., via an appropriate mediation device to enhance the user experience. That is, in addition to the audiovisual content of, for example, a movie, the user would perceive other effects, giving her/him the sensation of being part of the particular media, which should result in a worthwhile, informative user experience.

The current syntax and semantics of SEDL are specified in ISO [2011]. However, in this article, we provide an EBNF (Extended Backus–Naur Form)-like overview of SEDL.

```
SEM ::= [autoExtraction][DescriptionMetadata]
        (Declarations|GroupOfEffects|Effect|ReferenceEffect)+
```

SEM is the root element. It may contain an optional *autoExtraction* and *DescriptionMetadata* attribute followed by a sequence of *Declarations, GroupOfEffects, Effect*, and *ReferenceEffect* elements. The *autoExtraction* attribute is used to signal whether automatic extraction of a sensory effect from the media resource is preferable. The *DescriptionMetadata* attribute provides information about the SEM itself (e.g., authoring information) and aliases for classification schemes (CS) used throughout the whole description. The MPEG-7 description scheme [Manjunath et al. 2002] is used.

```
Declarations ::= (GroupOfEffects|Effect|Parameter)+
```

The *Declarations* element defines a set of SEDL elements, without instantiating them, for later use in an SEM via an internal reference. In particular, the *Parameter* may be used to define common settings used by several sensory effects similar to variables in programming languages.

A *GroupOfEffects* starts with a *timestamp* that provides information about the point in time when this group of effects should become available for the application. This information can be used for rendering purposes and synchronization with the associated media resource. XML streaming instructions as defined in MPEG-21 Digital Item Adaptation [Vetro and Timmerer 2005], have been adopted for this functionality. Furthermore, a *GroupOfEffects* shall contain at least two *EffectDefinition* for which no time stamps are required, as they are provided within the enclosing element. The actual *EffectDefinition* comprises all information pertaining to a single sensory effect.

```
Effect ::= timestamp EffectDefinition
```

An *Effect* is used to describe a single effect with an associated *timestamp*.

```
EffectDefinition  ::= [SupplementalInformation][activate][duration]
                      [fade-in][fade-out][alt][priority][intensity][position]
                      [adaptability][autoExtraction]
```

An *EffectDefinition* may have a *SupplementalInformation* element for defining a reference region from which the effect information may be extracted in case autoExtraction is enabled. Furthermore, several optional attributes are defined, which are defined as follows: *activate* describes whether the effect shall be activated; *duration* describes how long the effect shall be activated; *fade-in* and *fade-out* provide means for fading in/out effects, respectively; *alt* describes an alternative effect identified by a uniform resource identifier URI (e.g., in case the original effect cannot be processed); *priority* describes the priority of effects with respect to other effects in the same group of effects; *intensity* indicates the strength of the effect in percentage according to a predefined scale/unit (e.g., for wind the Beaufort scale is used); *position* describes the position from where the effect is expected to be received from the user's perspective (i.e., a three-dimensional space is defined in the standard); *adaptability* attributes enable the description of the preferred type of adaptation with a given upper and lower bound; *autoExtraction* with the same semantics as previously but only for a certain effect.

4.2.3. Sensory Effect Vocabulary (SEV). The Sensory Effect Vocabulary (SEV) defines a clear set of actual sensory effects to be used with the Sensory Effect Description Language (SEDL) in an extensible and flexible way. That is, it can be easily extended with new effects or by derivation of existing effects thanks to the extensibility feature of XML Schema. Furthermore, the effects are defined in a way to abstract from the authors intention and be independent from the end user's device setting. The sensory effect metadata elements or data types are mapped to commands that control sensory devices based on their capabilities. This mapping is usually provided by the media processing engine and deliberately not defined in this standard, that is, it is left open for industry competition. It is important to note that there is not necessarily a one-to-one mapping between elements or data types of the sensory effect metadata and sensory device capabilities. For example, the effect of hot/cold wind may be rendered on a single device with two capabilities, that is, a heater/air conditioner and a fan/ventilator. Currently, the standard defines the following effects.

Light, colored light, flash light for describing light effects with the intensity in terms of illumination expressed in [lux]. For the color information, a classification scheme (CS) is defined by the standard comprising a comprehensive list of common colors. Furthermore, it is possible to specify the color as RGB. The flash light effect extends the basic light effect by the frequency of the flickering in times per second.

Temperature describes a temperature effect of heating/cooling with respect to the Celsius scale. *Wind* provides a wind effect where it is possible to define its strength with respect to the Beaufort scale. *Vibration* allows one to describe a vibration effect with strength specified using a Richter magnitude scale. For the *water sprayer*, *scent*, and *fog* effect, the intensity is provided in terms of ml/h.

Finally, the *color correction* effect defines parameters that may be used to adjust the color information in a media resource to the capabilities of end-user devices. Furthermore, it is also possible to define a region of interest where the color correction shall be applied in case this is desirable (e.g., black/white movies with one additional color such as red).

5. QUALITY OF SERVICE, QUALITY OF EXPERIENCE, AND QUALITY OF SENSORY EXPERIENCE

5.1. Mulsemedia and Quality of Sensory Experience

New research perspectives on ambient intelligence are presented in Aarts and de Ruyter [2009], which includes also sensory experiences calling for a scientific framework to capture, measure, quantify, judge, and explain the user experience. In a previous paper [de Ruyter and Aarts 2004], the authors report on the effect additional light effects have on users. User studies showed that light effects are appreciated by users for both audio and visual contents.

In the context of the MPEG-V standardization ISO [2011], some work has been published related to sensory experience that is worth mentioning here. Suk et al. [2009] introduce a new generation of media service called Single Media Multiple Devices (SMMD) which is based on Sensory Effect Metadata (SEM) as defined in MPEG-V. In particular, the SMMD media controller is described that maps sensory effects on appropriate sensory devices for the proper rendering thereof. The main focus of this work is on implementation and engineering. An earlier version puts the controller in the context of Universal Plug and Play (UPnP), thus focusing also on implementation/engineering aspects [Pyo et al. 2008]. Yoon et al. [2010] present a framework for 4D broadcasting based on MPEG-V, that is, the main focus is on delivering additional representation formats in the MPEG-2 Transport Stream (M2TS) and its decoding within the home network environment including the actual service discovery. In this context, Waltl et al. [2013] provide an open-source end-to-end tool chain for creating and consuming multimedia content enriched with sensory effects compliant to MPEG-V based on off-the-shelf infrastructure.

Note that sensory effects are not limited to stationary installations, such as in home environments, as there is already research to bring sensory effects to mobile devices [Chang and O'Sullivan. 2005]. Furthermore, Kim et al. [2010] introduce, among others, new location-based mobile multimedia technology using ubiquitous sensor network-based five senses content. The temporal boundaries within which olfactory data can be used to enhance multimedia applications are investigated in Ademoye and Ghinea [2009], concluding that olfaction ahead of multimedia content is more tolerable than olfaction behind content.

Finally, Grega et al. [2008] provide a good overview of the state of the art in QoE evaluation for multimedia services with a focus on subjective evaluation methods which leads us to related work in the area of QoE models. Most of these models focus on a single modality (i.e., audio, image, or video only) or a simple combination of two modalities (i.e., audio and video). For the combination of audio and video content, one may employ the basic quality model for multimedia, as described in Hands [2004]. Another approach is known as the IQX hypothesis formulated as an exponential function [Hoßfeld et al. 2008]. In Pereira [2005] a triple-user characterization model for video adaptation and QoE evaluation is described that introduces at least three quality evaluation dimensions, namely, sensorial (e.g., sharpness, brightness), perceptual

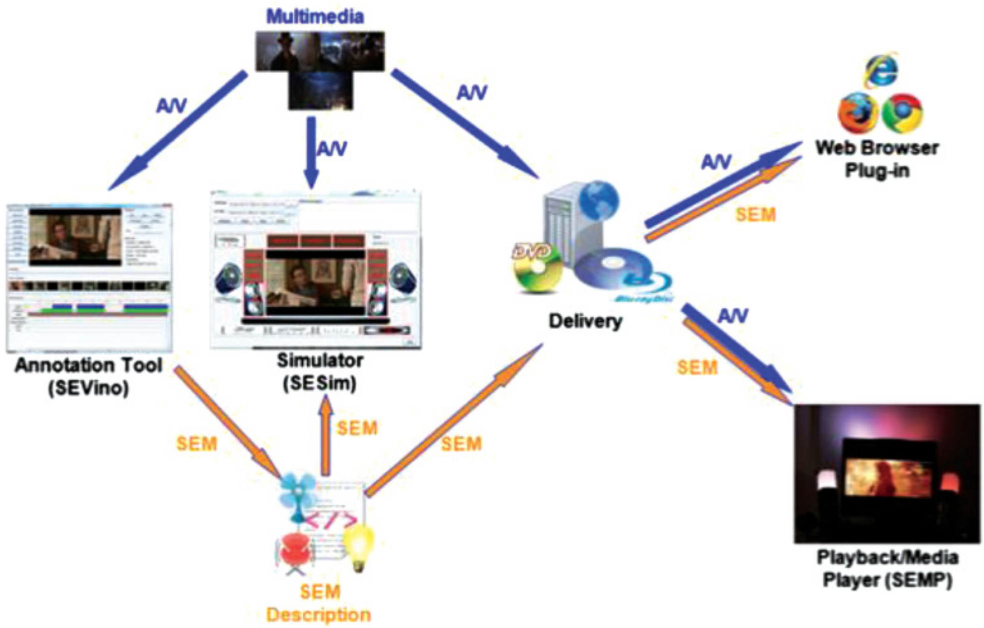

Fig. 3. Overview of end-to-end tool chain enabling to create, consume, and capture QuaSE.

(e.g., what/where is the content), and emotional (e.g., feeling, sensation) evaluation. Furthermore, it proposes adaptation techniques for the multimedia content and quality metrics associated to each of these layers. The focus is clearly on how an audio/visual resource is perceived, possibly taking into account certain user characteristics (e.g., handicaps) or natural environment conditions (e.g., illumination).

5.2. How to Create, Consume, and Capture QuaSE

In this section, we present a tool chain for creating and consuming media resources annotated with sensory effect, including means to capture the Quality of Sensory Experience (QuaSE). This set of tools is one of the first complete end-to-end tool chains offering an easy access from the generation of SEM descriptions till the consumption of audio/video (A/V) content accompanied by SEM descriptions in the context of the World Wide Web or the local playback devices.

Figure 3 illustrates the whole tool chain starting from the annotation tool (SEVino) on the left side. This tool receives the multimedia content for annotation with sensory effects and outputs the corresponding SEM description. These two assets can then be loaded into the simulator (SESim) located in the center of the figure or delivered via DVD, Blu-Ray, or the Internet. If the content is embedded into a website, the Web browser plug-in can playback the multimedia content within the Web browser and use the SEM description to steer appropriate devices. If the content is available on other means (e.g., DVD, Blu-Ray) then the stand-alone multimedia player (SEMP) can be used for enhancing the viewing experience. Note that the playback of the Web browser plug-in is performed by the Web browser itself. All tools are freely available under an open-source license and can be downloaded from the website of the Sensory Experience Lab (SELab) (`http://selab.itec.aau.at`).

The Sensory Effect Video Annotation (SEVino) tool allows for annotating video sequences with various sensory effects (e.g., wind, vibration, light) and generating

MPEG-V-compliant SEM descriptions. It is written in Java and for the actual decoding and rendering of the A/V files, the Java bindings for VLC[1] are used. Thus, it provides means for embedding the VLC player into a Java application and thus enables an application to support a lot of different codecs (e.g., H.264, MPEG-2) and file formats (e.g., MP4, AVI).

The Sensory Effect Simulator (SESim) allows for simulating sensory effects that are contained in SEM descriptions. The Sensory Effect Media Player (SEMP) is a DirectShow-based media player which supports the following devices for rendering sensory effects: the Philips amBX system (with two fans, a wall washer, two light-speakers, a subwoofer, and a wrist rumbler)[2], the Cyborg Gaming Lights (incl. high-power LEDs)[3], and the Vortex Activ device (comprises four slots for providing four different scents).[4] Note that as the media player uses DirectShow for playback, the media player can handle all formats and codecs which are supported either natively by Windows or via various codec packs.

Finally, the Web browser plugin is based on the AmbientLib which enables arbitrary applications to enrich the user experience with sensory effects. Thus, the library can be seen as an adaptation and processing engine between the virtual description of sensory effects and real devices capable of rendering the described effects. In particular, it provides functionalities to parse SEM descriptions, according to the MPEG-V standard, color calculation of video frames, and enables rendering of sensory effects on a variety of devices. AmbientLib provides an Application Programming Interface (API) and a Driver Interface (DI). The API enables embedding the library within any application and the DI is used for an easy integration of external devices (e.g., those supported also be SEMP) rendering sensory effects. One such application is the Web browser which allows the use of sensory effects with embedded video content on the World Wide Web such as YouTube.

In order to capture the Quality of Experience (QoE) enabled by mulsemedia, comprising traditional audiovisual content enriched with sensory effects, appropriate subjective quality assessments need to be conducted. Therefore, Waltl et al. [2012] provide a sensory effect dataset and test setups based on the open-source tools previously introduced. The test setups are aligned with ITU-T's recommendations for subjective quality assessments which provide the basis for studying the impact on the QoE when consuming multimedia assets annotated with sensory effects. Timmerer et al. [2012] describe the results of three subjective quality assessments in this domain based on methods defined by ITU-T P.910 and P.911, respectively [ITU-T 2008a, 2009b]. The main conclusions from these user studies are that genres such as action, sports, and also documenties benefit from additional sensory effects, while the impact on the QoE for genres like commercials and specifically news is not that much appreciated. Additionally, media resources with sensory effects may successfully mask visual quality degradations of the actual video content. In the extreme case, the low-quality version of the video enhanced with sensory effects receives higher ratings (on a mean opinion score scale) than the high-quality version of the video with sensory effects. Finally, in Rainer et al. [2012], the impact on the emotional state is investigated across different sites in Austria and Australia. The results indicate that the intensity of active emotions (e.g., interest, surprise, fun) are increased for video sequences with sensory effects compared to those without sensory effects. The results of the Austrian site also suggest that the intensity of passive emotions (e.g., worry, fear, anger) are decreased

[1] `http://www.videolan.org/vlc/`.(Last accessed: March 2014).
[2] `http://www.ambx.com/`.(Last accessed: March 2014).
[3] `http://www.cyborggaming.com/prod/ambx.htm`.(Last accessed: March 2014).
[4] `http://www.daleair.com/vortex-activ`.(Last accessed: March 2014).

for video sequences with sensory effects (compared to those without sensory effects), but with the results from the other sites, it does not yet allow for a general conclusion on whether passive emotions are decreased or increased in their intensity.

Finally, the ultimate goal is to define a utility model which tries to estimate the QoE of multimedia content enhanced with sensory effects based on various influence factors and features (See [Le Callet et al. 2013] for a general definition of QoE). These influence factors and features result from the QoE of the actual multimedia content and the QoE contributions of the individual sensory effects and the combinations thereof. The former can be estimated based on existing models (e.g., such as those referenced in the related work section), whereas the QoE contributions of the sensory effects, both individual and combinations, require further subjective quality assessments. Therefore, the results of such studies [Waltl et al. 2012; Timmerer et al. 2012] indicate a linear relationship between the number of effects and the actual QoE. Thus, the QoE of multimedia content enhanced with sensory effects is referred to as Quality of Sensory Experience (QuaSE) and can be estimated from the QoE of the audiovisual content without sensory effects ($\mathrm{QoE_{av}}$), as depicted as

$$\mathrm{QuaSE} := \mathrm{QoE_{av}}\left(\delta + \sum \mathrm{w_i b_i}\right).$$

In this utility model, w_i represents the weighting factor for a single sensory effect of type i (i.e., with the given setup as previously described, $i \in \{light(l), wind(w), vibration(v)\}$). Additional sensory effect types such as scent may be incorporated easily, for example, as soon as appropriate devices become available. The variables $b_i \in 0, 1$ depict the binary variables for each effect and are used to indicate whether an effect is present for a given setup. Finally, δ is used for fine-tuning an instantiation of the model.

6. RESEARCH CHALLENGES AND OPEN ISSUES

Mulsemedia is an emerging and exciting research area that we believe would extract much effort from the related academic and industrial communities. We have pointed out the challenges and possible research work in the Online Appendix after existing basic technical approaches and computational models are discussed. In this section, we will highlight R&D possibilities for the near future in order to further advance the technology, applications, and services based upon our understanding and project experience in the related fields. Technical advancement is expected to be made in and facilitated by effective algorithm development, substantial database building, meaningful applications, and wider user acceptance.

6.1. Mulsemedia – A Solution in Search of a Killer App?

6.1.1. Taste – The Last Frontier? For computation modeling of the functioning of human senses, as discussed in Section 3.1, most work has been done for audition and vision; significant recent interests have appeared toward olfaction and taction, and gustation is obviously the least investigated topic so far. We expect increasing activities to happen for gustation and the related issues. One challenge that we see is that, since taste buds are located in the mouth, devices that transmit sensations of taste will necessarily be invasive; alternatively, given the close relationship between taste and smell, it would also be interesting to monitor if the solution ultimately adopted will be to use (non-invasive) olfactory inputs to stimulate and engage gustation.

6.1.2. Attention Modeling. Human attention refers to the cognitive process of selectively concentrating on one aspect of the environment while ignoring other things [Anderson 2004]. As described in Section 2, inputs from one sense or different senses compete for human attention. Attention modeling has been formulated as the allocation of processing resources in humans, with a large number of examples in the visual sense [Itti et al. 1998; Zhang and Lin 2013] and joint audiovisual senses [Ma et al. 2005; You et al. 2007]. A comprehensive attention model should evaluate stimuli from all five senses, and this represents a meaningful research challenge for QoE exploration.

6.1.3. Building Databases. Appropriate databases play important roles in discovering necessary insights for modeling, model parameter determination, and model verification, as evidenced in the related existing visual and audio modeling [Campbell et al. 2009; Lin and Kuo 2011; Möller et al. 2011], and cross-database evaluation is essential toward model generality [Narwaria and Lin 2012; Narwaria et al. 2012]. There have been only a very limited number of databases available for odor (`http://www.odour.org.uk/information.html`, `http://senselab.med.yale.edu/odordb/?db$=$5`) and touch (`http://brl.ee.washington.edu/HapticsArchive/exp001.html`); more public databases are needed for mulsemedia (including gustation).

6.1.4. Mulsemedia and Performing Arts/Entertainment. 4D (and 5D) theatres are a staple attraction of theme parks worldwide and have been imparting 'novel' mulsemedia experiences to their visitors for some years now. The challenge will be to move such experiences from the theme parks into the mainstream. To some extent this is already happening: vibrating gaming chairs with integrated subwoofers (`http://www.4gamers.net/products/ps3/interactive-gaming-chair`), which make users 'feel' the action (and the bass in the audio) are gaining in popularity and becoming more affordable. Nonetheless, in order for mulsemedia to proliferate in these domains, we need to better understand how audiences react to mulsemedia effects; this will also enable script authors to effectively integrate them in the respective story lines.

6.1.5. Mulsemedia Integration, Synchronization, and Intensities. Effective integration of mulsemedia effects requires several questions to be answered: What mulsemedia combinations work in practice? In what doses/intensities? What synchronization requirements do new media such as olfactory and gustatory media need to satisfy in relation to their counterparts? These are all yet unanswered questions, which future research needs to target. Once clarified, new mulsemedia authoring tools would need to be written.

6.1.6. Wearable Mulsemedia. The miniaturization of sensors and computing devices alike has led to an increased focus on the potential of wearable technology: recently, both Google (through the Google Glass project, `http://www.google.com/glass/start/`) and Sony (through the SmartWig project, `http://www.bbc.co.uk/news/technology-25099262`) have brought to market wearable computing gadgets. If one thinks that individuals already 'wear' perfume and receive vibrating alerts when their smartphones are in silent mode, the potential of wearable devices to transmit mulsemedia content becomes obvious. Research will need to be done in order to understand how best to integrate such content in wearable devices, and indeed, how best to design such devices so that they can be purveyors of mulsemedia.

6.1.7. Mulsemedia and e-Learning. Mulsemedia authoring tools would also come in handy for e-learning systems. This, as e-learning systems stand to gain potential benefits from olfaction-enhanced mulsemedia applications (for instance), as the online learning of certain subject matters (e.g., chemistry) may be further enhanced by the addition of the corresponding smells if it were possible to transmit odors, or more precisely, transmit commands to a smell-generating device to mix and emit the required scent over the Internet. Such future work would of necessity need to explore in what contexts and to which extent mulsemedia improves communications. In so doing, guidelines about how exactly to use mulsemedia to achieve a more accurate knowledge transfer would need to be elaborated.

6.1.8. Mulsemedia and e-Commerce. The options to feel the texture of a shirt that one wishes to buy, to smell the fragrance one is contemplating of purchasing, of inhaling the aroma, as well as seeing, tasting and experiencing the texture of a gourmet dish before booking a table at the restaurant serving it, all have the potential of moving from the realm of possibilities to that of reality. In so doing, the touch/taste/smell barriers currently characteristic of e-Commerce will be overcome.

6.1.9. User Acceptance and Experience. We started off this section by highlighting the need for a mulsemedia killer app. Whilst we have detailed, among others, what we believe to be potentially interesting mulsemedia developments, we cannot make any predictions for what a killer mulsemedia app might be. One thing, however, is certain: user acceptance, and more importantly, take-up is essential for any killer app. In order to do this, future work needs to undertake mulsemedia QoE studies to better understand how mulsemedia users react to such experiences. Moreover, in so doing, such efforts would also inform the development of objective mulsemedia QoE metrics.

6.2. Final Thought

"Seeing is believing" is an often-quoted idiom. Perhaps not so well known is the fact that the complete idiom, as penned by its author, the 17th century English clergyman, Thomas Fuller, is actually "Seeing is believing, but feeling is the truth." We subscribe to this statement, but feel that, for mulsemedia, the idiom is (at least) three sentences too short.

REFERENCES

E. Aarts and B. de Ruyter. 2009. New research perspectives on ambient intelligence. *J. Ambient Intell. Smart Environ.* 1, 1, 5–14.

O. Ademoye and G. Ghinea. 2009. Synchronization of olfaction-enhanced multimedia. *IEEE Trans. Multimedia* 11, 3, 561–565.

J. R. Anderson. 2004. *Cognitive Psychology and Its Implications* 6th Ed. Worth Publishers.

J. G. Apostolopoulos, P. A. Chou, B. Culbertson, T. Kalker, M. D. Trott, and S. Wee. 2012. The road to immersive communication. *Proc. IEEE* 100, 4, 974–990.

S. Ayabe–Kanamura, I. Schicker, M. Laska, R. Hudson, H. Distel, T. Koboyakawa, and S. Saito. 1998. A Japanese-German cross-cultural study. *Chem. Sens.*, 23, 1, 31–38.

A. Bodnar, R. Corbett, and D. Nekrasovski. 2004. AROMA: Ambient awareness through olfaction in a messaging application: Does olfactory notification make 'scents'? In *Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI'04).* 183–190.

S. Boyd-Davis, G. Davies, R. Haddad, and M. Lai. 2006. Smell Me: Engaging with an interactive olfactory game. In *Proceedings 25th Annual Meeting of the Human Factors and Ergonomics Society*, 25–40.

S. A. Brewster, D. K. Mcgookin, and C. A. Miller. 2006. Olfoto: Designing a smell-based interaction. In *Proceedings of the Conference on Human Factors in Computing Systems*. 653–662.

D. Campbell, E. Jones, and M. Glavin. 2009. Audio quality assessment techniques—A review, and recent developments. *Signal Process.*, 89, 8, 1489–1500.

C.-C. Carbon and M. Jakesch. 2013. A model for haptic aesthetic processing and its implications for design. *Proc. IEEE*. 101, 9, 2123–2133.

J. P. Cater. 1992. The nose have it! Letters to the editor. *Presence* 1, 4, 493–494.

A. Chang and C. O'Sullivan. 2005. Audio-haptic feedback in mobile phones. In *Proceedings of the Extended Abstracts on Human Factors in Computing Systems (CHI EA'05)*. ACM, New York, NY, 2005, 1264–1267.

A. D. Craig. 2003. Interoception: The sense of the physiological condition of the body. *Curr. Opinion Neurobiol*. 13, 4, 500–505.

A. R. Damasio. 1989. Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition* 33, 1–2, 25–62.

B. de Ruyter and E. Aarts. 2004. Ambient intelligence: Visualizing the future. In *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI'04)*. ACM Press, New York, NY, 203–208.

P. Dimaggio. 1997. Culture and cognition. *Ann. Rev. Sociol.* 23, 263–287.

H. Q. Dinh, N. Walker, L. F. Hodges, C. Song, and A. Kobayashi. 1999. Evaluating the importance of multi-sensory input on memory and the sense of presence in virtual environments. In *Proceedings of the Virtual Reality Annual International Symposium*. 222–228.

C. Fadel and C. Lemke. 2008. Multimodal learning through media: What the research says. CISCO Systems Report.

G. Ghinea and O. Ademoye. 2010a. Perceived synchronization of olfactory multimedia. *IEEE Trans. Syst. Man Cybernet.-Part A* 40, 4, 657–663.

G. Ghinea and O. Ademoye. 2010b. A user perspective of olfaction-enhanced mulsemedia. In *Proceedings of the International Conference on Management of Emergent Digital EcoSystems (MEDES'10)*. 277–280.

G. Ghinea and O. Ademoye. 2011. Olfaction-enhanced multimedia: Perspectives and challenges. *Multimedia Tools Appl.* 55, 3, 601–626.

G. Ghinea and O. Ademoye. 2012. The sweet smell of success: Enhancing multimedia applications with olfaction. *ACM Trans. Multimedia Comput. Commun. Appl.* 8, 1, 2.

E. B. Goldstein. 2013. *Sensation and Perception*. Cengage Learning.

R. Gray, C. Spence, C. Ho, and H. Z. Tan. 2013. Efficient multimodal cuing of spatial attention *Proc. IEEE* 101, 9, 2113–2121.

M. Grega, L. Janowski, M. Leszczuk, P. Romaniak, and Z. Papir. 2008. Quality of experience evaluation for multimedia services - Szacowanie postrzeganej jako sci uslug (QoE) komunikacji multimedialnej. *Przegla d Telekomunika- cyjny* 81, 4, 142–153.

S. Gumtau. 2011. Affordances of touch in multisensory embodied interface design. Ph.D., Dissertation University of Portsmouth, U.K.

D. Hands. 2004. A basic multimedia quality model. *IEEE Trans. Multimedia* 6, 6, 806–816.

M. L. Heilig. 1962. Sensorama simulator. United States Patent 3,050,870. Filed January 10, 1961, Patented August 28, 1962.

P. Hinterseer and E. Steinbach. 2006. A psychophysically motivated compression approach for 3D haptic data. In *Proceedings of the 14th Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*. 35–41.

C. Ho and C. Spence. 2005. Olfactory facilitation of dual-task performance. *Neurosci. Lett.* 389, 1, 35–40.

T. Hoßfeld, D. Hock, P. Tran-Gia, K. Tutschku, and M. Fiedler. 2008. Testing the IQX hypothesis for exponential interdependency between QoS and QoE of voice codecs iLBC and G.711. In *Proceedings of the 18th ITC Specialist Seminar on Quality of Experience*.

Y. Ishibashi, T. Kanbara, and S. Tasaka. 2004. Inter-stream synchronization between haptic media and voice in collaborative virtual environments. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*. ACM, New York, NY, 604–611.

L. Itti, C. Koch, and E. Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Patt. Anal. Mach. Intell.* 20, 11, 1254-9.

ITU. 2008a. Subjective video quality assessment methods for multimedia applications. ITU-T Rec. P.910.

ITU. 2008b. Subjective audiovisual quality assessment methods for multimedia applications. ITU-T Rec. P.911.

R. Jain. 2003. Experiential computing. *Commun. ACM* 46, 7, 48–55.

N. Jayant, J. Johnston, and R. Safranek. 1993. Signal compression based on models of human perception. *Proc. IEEE* 81, 10, 1385–1422.

L. Jones, C. A. Bowers, D. Washburn, A. Cortes, and R. V. Satya. 2004. The effect of olfaction on immersion into virtual environments. In *Human Performance, Situation Awareness and Automation: Issues and Considerations for the 21st Century*. Lawrence Erlbaum Associates, 282–285.

K. Kahol, P. Tripathi, T. Mcdaniel, L. Bratton, and S. Panchanathan. 2006. Modeling context in haptic perception, rendering, and visualization. *ACM Trans. Multimedia Comput. Commun. Appl.* 2, 3, 219–240.

J. Kammerl, I. Vittorias, V. Nitsch, B. Faerber, E. Steinbach, and S. Hirche. 2010. Perception-based data reduction for haptic force-feedback signals using adaptive deadbands. *Presence, Teleoper. Virtual Environ.* 19, 5, 450–462.

D. Kahneman. 2003. A perspective on judgement and choice. *Am. Psychol.* 58, 9, 697–720.

J. N. Kaye. 2001. Symbolic olfactory display. Master's Thesis, Massachusetts Institute of Technology, MA. http://www.media.mit.edu/ jofish/thesis/.

H. Kim, H.-J. Kwon, and K.-S. Hong. 2010. Location awareness-based intelligent multi-agent technology. *Multimedia Syst*. 16, (4–5), 275–292.

R. L. Klatzky, D. Pawluk, and A. Peer. 2013. Haptic perception of material properties and implications for applications. *Proc. IEEE* 101, 9, 2081–2092.

P. Le Callet, S. Möller, and A. Perkis. (Eds). 2013. Qualinet white paper on definitions of quality of experience. White paper, European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003). Lausanne, Switzerland, Version 1.2.

W. Lin. 2006. Computational models for just-noticeable difference. In *Digital Video Image Quality and Perceptual Coding*, H. R. Wu and K. R. Rao, (Eds.), CRC Press, Chapter 9.

W. Lin and C.-C. Jay Kuo. 2011. Perceptual visual quality metrics: A survey. *J. Visual Commun. Image Represent.* 22, 4, 297–312.

K. Liu and S. R. Gulliver. 2013. Semiotics in building space for working and living. In *Intelligent Building: Design, Management and Operation,* D. Clements-Croome, (Ed.), ICE Publishing.

Z. Lu, W. Lin, X. Yang, E. Ong, and S. Yao. 2005. Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation. *IEEE Trans. Image Process.*, 14, 11, 1928–1942.

Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang. 2005. A generic framework of user attention model and its application in video summarization. *IEEE Trans. Multimedia*. 7, 5, 907–919.

R. E. Mayer. 2003. Elements of a science of e-learning. *J. Edu. Comput. Res.* 29, 3, 297–313.

B. S. Manjunath, P. Salembier, and T. Sikora. 2002. *Introduction to MPEG-7: Multimedia Content Description Interface*. John Wiley 8. Sons Inc.

R. Marois and J. Ivanoff. 2005. Capacity limits of information processing in the brain. *Trends Cognitive Sci.*, 9, 6, 296–305.

T. Metzinger. 1995. Faster than thought: Holism, homogeneity and temporal coding. In *Conscious Experience* T. Metzinger (Ed.), Imprint Academic, 425–461.

A. Mochizuki, T. Amada, S. Sawa, T. Takeda, S. Motoyashiki, K. Kohyama, M. Imura, and K. Chihara. 2004. Fragra: A visual-olfactory VR game. In *Proceedings of the ACM SIGGRAPH (SIGGRAPH'04)*. ACM Press, New York, NY, 123.

S. Möller, W.-Y. Chan, N. Côté, T. H. Falk, A. Raake, and M. Wältermann. 2011. Speech quality estimation: Models and trends. *IEEE Signal Process. Mag.* 28, 6, 18–28.

G. Morrot, F. Brochet, and D. Dubourdieu. 2001. The color of odors. *Brain Lang.* 79, 2, 309–320.

M. Narwaria and W. Lin. 2012. SVD-based quality metric for image and video using machine learning. *IEEE Trans. Syst. Man Cybernet. Part B* 42, 2, 347–364.

M. Narwaria, W. Lin, I. Mcloughlin, S. Emmanue, and L. T. Chia. 2012. Nonintrusive quality assessment of noise suppressed speech with mel-filtered energies and support vector regression. *IEEE Trans. Audio Speech Lang. Process.* 20, 4, 1217–1232.

T. Nakamoto, S. Otaguro, M. Kinoshita, M. Nagahama, K. Ohinishi, and T. Ishida. 2008. Cooking up an interactive olfactory game display. *IEEE Comput. Graph. Appl.* 28, 1, 75–78.

H.-C. Nothdurft. 2000. Salience from feature contrast: Additivity across dimensions. *Vis. Res.* 40, 10–12, 1183–1201.

M. A. Otaduy, C. Garre, and M. C. Lin. 2013. Representations and algorithms for force-feedback display. *Proc. IEEE* 101, 9, 2068–2080.

F. Pereira. 2005. A triple user characterization model for video adaptation and quality of experience evaluation. In *Proceedings of the 7th IEEE Workshop on Multimedia Signal Processing*. 1–4.

S. Pyo, S. Joo, B. Choi, M. Kim, and J. Kim. 2008. A metadata schema design on representation of sensory effect information for sensible media and its service framework using UPnP. In *Proceedings of the 10th International Conference on Advanced Communication Technology (ICACT'08)*. 1129–1134.

B. Rainer, M. Waltl, 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 *Proceedings of the 4th International Workshop on Quality of Multimedia Experience (QoMEX'12)*. IEEE, 278–283.

E. Reinhard, A. A. Efros, J. Kautz, and H.-P. Seidel. 2013. On visual realism of synthesized imagery. *Proc. IEEE* 101, 9, 1998–2007.

A. Revonsuo. 1999. Binding and the phenomenal unity of consciousness. *Consciou. Cognition* 8, 2, 173–185.

G. Richard, S. Sundaram, and S. Narayanan. 2013. An overview on perceptually motivated audio indexing and classification. *Proc. IEEE* 101, 9, 1939–1954.

L. A. Rowe and R. Jain. 2005. ACM SIGMM retreat report on future directions in multimedia research. *ACM Trans. Multimedia Comput. Commun. Appl.* 1, 1, 3–13.

J. S. Rubinstein, D. E. Meyer, and J. E. Evans. 2001. Executive control of cognitive processes in task switching. *J. Exp. Psychol. Human Percep. Perform.* 27, 4, 763.

N. Sarter. 2013. Multimodal support for interruption management: Models, empirical findings, and design recommendations. *Proc. IEEE* 101, 9, 2105–2112.

P. H. Schiller. 1986. The central visual system. *Vision Res.* 26, 9, 1351–1386.

C. Seungmoon and K. J. Kuchenbecker. 2013. Vibrotactile display: Perception, technology, and applications. *Proc. IEEE* 101, 9, 2093–2104.

J. R. Smythies. 1994a. *The Walls of Plato's cave*. Avebury.

J. R. Smythies. 1994b. Requiem for the identity theory. *Inquiry* 37, 311–329.

R. K. Stamper. 1973. *Information in Business and Administrative Systems*. John Wiley & Sons, New York.

E. Steinbach, S. Hirche, M. Ernst, F. Brandi, R. Chaudhari, J. Kammerl, and I. Vittorias. 2012. Haptic communications. *Proc. IEEE* 100, 4, 937–956.

C. B. Suk, J. S. Hyun, and L. H. Yong. 2009. Sensory effect metadata for SMMD media service. In *Proceedings of the 4th International Conference on Internet and Web Applications and Services*. IEEE Computer Society, 649–654.

ISO. 2011. ISO/IEC 23005-3 FDIS Information technology—Media context and control—Part 3: Sensory information. ISO.

C. Timmerer, M. Waltl, B. Rainer, and H. Hellwagner. 2012. Assessing the quality of sensory experience for multimedia presentations. *Signal Process. Image Commun.* 27, 8, 909–916.

R. Tortell, D. P. Luigi, A. Dozois, S. Bouchard, J. F. Morie, and D. Ilan. 2007. The effects of scent and game play experience on memory of a virtual environment. *VirtualReality* 11, 1, 61–68.

A. Vetro and C. Timmerer. 2005. Digital item adaptation: Overview of standardization and research activities. *IEEE Trans. Multimedia* 7, 3, 418–426.

M. Waltl, C. Timmerer, B. Rainer, and H. Hellwagner. 2012. Sensory effect dataset and test setups. In *Proceedings of the 4th International Workshop on Quality of Multimedia Experience (QoMEX'12)*. IEEE, 115–120.

M. Waltl, B. Rainer, C. Timmerer, and H. Hellwagner. 2013. An end-to-end tool chain for sensory experience based on MPEG-V *Signal Process. Image Commun.* 28, 2, 136–150.

A. Williams, S. Langron, and A. Noble. 1984. Influence of appearance on the assessment of aroma in Bordeaux wines by trained assessors. *J. Institute Brew.* 90, 250–253.

H. R. Wu, A. Reibman, W. Lin, F. Pereira, and S. S. Hemami. 2013. Perceptual visual signal compression and transmission. *Proc. IEEE* 101, 9, 2025–2043.

X. Yang, W. Lin, Z. Lu, E. Ong, and S. Yao. 2005. Just noticeable distortion model and its applications in video coding. *Signal Process. Image Commun.* 20, 7, 662–680.

A. L. Yarbus. 1967. Eye movements during perception of complex objects. In *Eye Movements and Vision*. Springer, 171–211.

A. Yazdani, E. Kroupi, J. Vesin, and T. Ebrahimi. 2012. Electroencephalogram alterations during perception of pleasant and unpleasant odors. In *Proceedings of the 4th International Workshop on Quality of Multimedia Experience (QoMEX'12)*. IEEE, 272–277.

K. Yoon, B. Choi, E.-S. Lee, and T.-B. Lim. 2010. 4-D Broadcasting with MPEG-V. In *Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP)*. 257–262.

J. You, U. Reiter, M. M. Hannuksela, M. Gabbouj, and A. Perkis. 2010. Perceptual-based quality assessment for audio-visual services: A survey. *Signal Process. Image Commun.* 25, 7, 482–501.

J. You, G. Liu, L. Sun, and H. Li. 2007. A multiple visual models based perceptive analysis framework for multilevel video summarization. *IEEE Trans. Circuits Syst. Video Technol.* 17, 3, 273–285.

W. A. Yost and D. W. Nielsen. 1985. *Fundamentals of Hearing*. Holt, Rinehart and Winston, New York.

L. M. Zhang and W. Lin. 2013. *Selective Visual Attention: Computational Models and Applications.* John Wiley & Sons.

Received December 2013; revised March 2014; accepted April 2014