Immersive Multisensory Digital Twins: concept, methods and case study

Charles Javerliat, Pierre Raimbaud, Pierre-Philippe Elst, Eliott Zimmermann, Sophie Villenave, Martin Guesney, Mylène Pardoen, Patrick Baert, Guillaume Lavoué

Immersive Multisensory Digital Twins: concept, methods and case study

CHARLES JAVERLIAT, PIERRE RAIMBAUD, PIERRE-PHILIPPE ELST, and ELIOTT ZIMMERMANN, Centrale Lyon, Univ Lyon, CNRS, INSA Lyon, UCBL, LIRIS, UMR5205, ENISE, France
SOPHIE VILLENAVE, Vassiléo, Montpellier, France
MARTIN GUESNEY and MYLÈNE PARDOEN, Maison des sciences de l'Homme Lyon Saint-Étienne, France
PATRICK BAERT and GUILLAUME LAVOUÉ, Centrale Lyon, Univ Lyon, CNRS, INSA Lyon, UCBL, LIRIS, UMR5205, ENISE, France

This paper presents the concept of multisensory digital twins, a brief state of the art of existing techniques for both capturing and immersive rendering of multisensory stimuli, and a case study of multisensory digital twin creation for cultural heritage.

1 INTRODUCTION

Virtual reality (VR) refers to the set of technologies that allows a user to be transported into a world artificially generated by a computer – also called virtual environment (VE). The user has the ability to interact in this environment, e.g., to move around, select and manipulate objects [2]. This ability to interact as well as the stimulation of several of the user's senses, allows one to *be present* in a VE as if it was a real environment. While most existing VR experiences are limited to audiovisual stimuli, recent surveys suggest going beyond this unique kind of stimulus. These notably show that the inclusion of auditory, olfactory, and tactile (including thermoception) stimuli seems to significantly enhance VR user experience, for several aspects [12, 13]. A notable one is the benefits that come with increased degrees of presence and immersion for VR users. In addition, more and more affordable techniques have been recently developed to capture, create and render multisensory stimuli, whose digital representation is referred to as mulsemedia [3, 6] – for *multiple sensorial media*. These techniques enable the creation of what we may call **multisensory digital twins**. This concept extends the original digital twin concept beyond 3D graphics and data, by including sensory information (odors, tactile sensations, taste, sounds...). Such virtual multisensory replicas open the way to reproducing true **sensitive experiences**. Its application for sensory reconstitution of past scenes, for example, opens up very interesting perspectives for museography, and the preservation and transmission of tangible and intangible cultural heritage. This paper presents a brief state-of-the-art of existing techniques to record and immersively render multisensory stimuli, and a case study of multisensory digital twin creation for cultural heritage.

2 TECHNIQUES FOR RECORDING AND RENDERING SENSORY STIMULI

Visual. To capture the visual appearance of a real scene, two solutions are possible. The first one is to manually create objects, e.g., from photographs using modeling software such as Blender, Maya, or 3DS Max. This method is costly in terms of time and human resources and can also introduce an artistic bias. For greater fidelity, and especially for cultural heritage where authenticity is crucial, it would be preferable to use a method that could limit or even prevent deviating from reality. In this regard, photogrammetry allows for the automatic capture of the appearance of an object using Multi-View Stereo [20] or Structure-from-Motion techniques [11, 19]. It allows for the reconstruction of a three-dimensional model, its color, and its physical properties, from a set of photographs, all with an accuracy of up to a millimeter.

However, the rendering of a visual experience is not only about presenting monuments, objects, or other fixed visual environments. The rendering of 3D movements is also crucial – especially human movements, to ensure the visual dynamics of the scene are credible. For the capture of 3D movements, motion capture methods can be used. Reflective markers are placed on actors, their joints, and the middle of their limbs, and captured by infrared cameras. As several of these cameras perceive the same points, a calculation by triangulation makes it possible to find the position of the markers in the tridimensional space. Once these data are acquired, they can be used to animate a digital model. Although offering a precision below the millimeter, this optical method is not without flaws. First, its precision is very sensitive to the capture environment. Operating in the infrared range, the sunlight, notably present outdoors, can interfere with the capture, therefore limiting the use of this technique to environments where conditions are controlled, such as in a film studio. Secondly, wearing a suit and markers can be annoying for the captured persons, and, even more relevant, can introduce a bias in their behavior. Another option, possible due to advances in the fields of machine learning and computer vision, is the reconstruction of 3D movements using video capture only (without markers). These new digital techniques, which can be found in the literature under the terms of Markerless 3D Human Pose Estimation, are a hot topic, with thousands of research articles on the subject every year. From a multitude of examples, neural networks can learn to detect points of interest, used to determine the position of body joints. These new methods are not yet perfect – considering their accuracy is currently in the centimeter range, but they open up attractive prospects for motion capture.

Audio. A large variety of recording techniques for capturing spatialized sounds, either in stereo (2 channels) or in surround (at least 5 channels) currently exist [4, 16]. For the case of immersive multisensory digital twins, the most relevant techniques are: i) the *binaural* capture, which consists of recording the sound as close as possible to the two eardrums of a person (called native binaural) through miniature microphones; ii) the capture with *localized microphones*, which takes into account the existing sources in an environment by assigning a microphone near a source, and allows to obtain spatialized information of the environment through reconstitution. For a 3DoFs virtual environment (e.g., 360 video), binaural capture is the most relevant one, whereas, for a 6DoFs environment (in which users can move), the best choice is the capture with localized microphones. This is particularly true since 3D game engines used for VR allow for the positioning of audio sources in a virtual world, and for the automatic rendering of spatialization in the audio system.

For sound restitution in VR, the stereo system (with headphones) is the most frequently used, since directly integrated into head-mounted displays – HMDs. However, more accurate technologies may also be used such as Wave Field Synthesis, a sound diffusion technology to recreate acoustic fields of one or more sources within a listening area delimited by the speaker system. Its main benefit is its large listening area that allows the listener to move around.

Olfaction. As raised by Ghinea et al. [5], olfaction has great potential in multisensory immersive experiences due to its influence on emotions, mood, and memory. Several authors proposed wearable devices compatible with HMDs; in particular, Javerliat et al. [7] proposed an open-source reproducible device compatible with autonomous HMDs (e.g., Meta Quest 2). Several forms of odorant conditioning exist: liquid, solid, and gaseous, which each involve different diffusion mechanisms. Gaseous form diffusion allows for great precision but usually involves complex technical devices; therefore, it should be highlighted that most wearable devices currently use the atomization of liquid phase odors.

The capture of odors can be done objectively through the use of passive or active olfactometers. These are generally based on optical or electrical processes allowing the determination of the chemical composition of the ambient air. Another approach is to rely on expert olfactory designers able to capture, identify and reproduce odors very precisely.

Haptics. As described by Muender et al. [15]: *humans perceive haptics through their cutaneous and kinesthetic systems enabling the perception of material characteristics of surfaces and objects as well as position and movement of their own body.* Nowadays, many technologies exist to render a wide range of haptic stimuli: pressure, vibration, surface shape, texture, temperature, forces, proprioception [15]. Regarding its capture, haptography [10] is a discipline dedicated to this. While the majority of virtual haptic sensations are programmed and manually parameterized following theoretical models, haptography seeks to empirically capture the sensations produced when an object is touched. Romano and Kuchenbecker [18] developed a prototype haptograph: a pen instrumented with velocity, acceleration, and force sensors.

In addition, it should be noticed that, among haptic perception, proprioception plays a particular role. This sense allows us to know the position of our limbs in space. It has been shown in the literature that this "sixth sense" strongly contributes to immersion in VEs, especially in terms of presence or engagement [1]. The accurate tracking and representation in the virtual environment of the participant body are crucial to improve embodiment [9].

Taste. Among the senses explored in VR, taste is among the last and least studied, notably for practical (difficulties for VR users to take and ingest food) and technical reasons (difficulties to artificially create taste sensations). Beyond real food ingestion, taste synthesis technologies have been tested outside of immersive VEs: electrophoresis [14], electrical stimulation [17], or tongue thermal stimulation [8]. However, the effects still remain far from the taste of real food.

3 A CASE STUDY FOR CULTURAL HERITAGE: AUDIO, VISUAL AND OLFACTIVE RECORDING

The principles of multisensory recording and rendering described above are illustrated in the PROMESS project (financed by Auvergne-Rhône-Alpes region in France), whose objective is the immersive multisensory restitution of scenes from the past. One of the targeted scenarios is the immersion of visitors in the construction site of Notre Dame de Paris cathedral in the Middle Ages. To ensure the technical and historical fidelity of craftsmen's workshops, sensory recordings are conducted on Guédelon castle[1], a worksite using Middle Ages techniques and materials.

Fig. 1. Audio-visual recording of a carpenter on the Guedelon site

Visual. Artisans' workshops are captured by photogrammetry using iPhone 13 Pro and Agisoft Metashape software[2]. In addition to the visual environment, artisans' 3D movements are also recorded through motion capture (using optical markers from OptiTrack system), to transfer motions to virtual avatars[3].

Audio. Each visual recording is accompanied by a simultaneous sound recording to allow for perfect synchronization between images and sound. Several directional microphones are deployed for isolated captures of each sound point of interest. An ultra-directional microphone on a pole allows to follow craftsmen's gestures as close as possible, without perturbations. Contact microphones, glued to the material being worked on (stone, wood, metal), allow for isolated sound capture, regardless of the surroundings. Oppositely, ambient microphones (stereophonic pair or ambient microphone) are used to capture the general atmosphere.

Olfaction. Craftsmen's workshops contain profession-related odors (e.g., wood, stone, lime). An expert olfactory designer is in charge of their capture and render.

[1] https://www.guedelon.fr/en

[2] Example available at https://youtu.be/69NUwaEnRLc

[3] Examples available at https://www.youtube.com/watch?v=1RasDCEJuSQ

Rendering. The immersive restitution of recorded sensory stimuli is planned through the use of a high quality VR headset such as the Varjo XR-3. For odors, the Nebula device [7] can be easily attached to it, allowing for smell diffusion close to the nose. For sound diffusion, two choices are considered: binaural headphones, and ambisonic diffusion via a set of speakers. This later would allow for a better fidelity of sound, but could limit the amplitude of VR users motions.

REFERENCES

[1] Ferran Argelaguet, Ludovic Hoyet, Michaël Trico, and Anatole Lécuyer. 2016. The role of interaction in virtual embodiment: Effects of the virtual hand representation. In *2016 IEEE virtual reality (VR)*. IEEE, IEEExplore, NY, USA, 3–10.

[2] Doug A Bowman, Ernst Kruijff, Joseph J LaViola, and Ivan Poupyrev. 2001. An introduction to 3-D user interface design. *Presence* 10, 1 (2001), 96–108.

[3] Alexandra Covaci, Longhao Zou, Irina Tal, Gabriel Miro Muntean, and Gheorghita Ghinea. 2018. Is multimedia multisensorial? - A review of mulsemedia systems. *Comput. Surveys* 51, 5 (2018), 23–27.

[4] Paul Geluso. 2021. 3D acoustic recording. In *3D Audio*. Routledge, London, UK, 228–255.

[5] Gheorghita Ghinea and Oluwakemi A Ademoye. 2011. Olfaction-enhanced multimedia: perspectives and challenges. *Multimedia Tools and Applications* 55 (2011), 601–626.

[6] 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 (oct 2014), 23–27.

[7] Charles Javerliat, Pierre-Philippe Elst, Anne-Lise Saive, Patrick Baert, and Guillaume Lavoué. 2022. Nebula: An Affordable Open-Source and Autonomous Olfactory Display for VR Headsets. In *Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology*. Association for Computing Machinery, New York, NY, United States, 1–8.

[8] Kasun Karunanayaka, Nurafiqah Johari, Surina Hariri, Hanis Camelia, Kevin Stanley Bielawski, and Adrian David Cheok. 2018. New thermal taste actuation technology for future multisensory virtual reality and internet. *IEEE transactions on visualization and computer graphics* 24, 4 (2018), 1496–1505.

[9] Konstantina Kilteni, Raphaela Groten, and Mel Slater. 2012. The sense of embodiment in virtual reality. *Presence: Teleoperators and Virtual Environments* 21, 4 (2012), 373–387.

[10] Katherine J. Kuchenbecker. 2008. Haptography: Capturing the Feel of Real Objects to Enable Authentic Haptic Rendering (Invited Paper). In *Proceedings of the 2008 Ambi-Sys Workshop on Haptic User Interfaces in Ambient Media Systems (HAS '08)*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Quebec City, Canada, Article 3, 3 pages.

[11] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. 2021. Pixel-perfect structure-from-motion with featuremetric refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEExplore, NY, USA, 5987–5997.

[12] Daniel Martin, Sandra Malpica, Diego Gutierrez, Belen Masia, and Ana Serrano. 2022. Multimodality in VR: A Survey. *Comput. Surveys* 54, 10s (sep 2022), 36.

[13] Miguel Melo, Guilherme Goncalves, Pedro Monteiro, Hugo Coelho, José Vasconcelos-Raposo, and Maximino Bessa. 2022. Do Multisensory Stimuli Benefit the Virtual Reality Experience? A Systematic Review. *IEEE Transactions on Visualization and Computer Graphics* 28, 2 (2022), 1428–1442.

[14] Homei Miyashita. 2021. TTTV (taste the TV): Taste presentation display for "licking the screen" using a rolling transparent sheet and a mixture of liquid sprays. In *Adjunct Proceedings of the 34th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, United States, 37–40.

[15] Thomas Muender, Michael Bonfert, Anke Verena Reinschluessel, Rainer Malaka, and Tanja Döring. 2022. Haptic fidelity framework: Defining the factors of realistic haptic feedback for virtual reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, United States, 1–17.

[16] Edwin Pfanzagl-Cardone. 2019. *The Art and Science of Surround-and Stereo-Recording*. Springer, Wien, Austria.

[17] Nimesha Ranasinghe, Thi Ngoc Tram Nguyen, Yan Liangkun, Lien-Ya Lin, David Tolley, and Ellen Yi-Luen Do. 2017. Vocktail: A virtual cocktail for pairing digital taste, smell, and color sensations. In *Proceedings of the 25th ACM international conference on Multimedia*. Association for Computing Machinery, New York, NY, United States, 1139–1147.

[18] Joseph M. Romano and Katherine J. Kuchenbecker. 2012. Creating Realistic Virtual Textures from Contact Acceleration Data. *IEEE Transactions on Haptics* 5, 2 (2012), 109–119. https://doi.org/10.1109/TOH.2011.38

[19] Johannes L Schonberger and Jan-Michael Frahm. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEExplore, NY, USA, 4104–4113.

[20] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. 2016. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, Springer, New York, NY, United States, 501–518.