A Classification Scheme for Multi-Sensory Augmented Reality

Robert W. Lindeman
Worcester Polytechnic Institute
gogo@wpi.edu

Haruo Noma
ATR International
noma@atr.jp

Abstract

We present a new classification framework for describing augmented reality (AR) applications based on where the mixing of real and computer-generated stimuli takes place. In addition to "classical" visual AR techniques, such as optical-see-through and video-see-through AR, our framework encompasses AR directed at the other senses as well. This "axis of mixing location" is a continuum ranging from the physical environment to the human brain. There are advantages and disadvantages of mixing at different points along the continuum, and while there is no "best" location, we present sample usage scenarios that illustrate the expressiveness of this classification approach.

CR Categories: H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems; Artificial, augmented, and virtual realities

Keywords: Augmented reality, video, audio, haptics, olfactory, gustatory.

1 Introduction

Augmented reality (AR) is the mixing of computer-generated (CG) or mediated stimuli with real-world (RW) stimuli. In the visual domain, example applications of AR techniques include annotation of live video streams for informational purposes [2, 3], instruction for performing maintenance tasks [3], and collaboration [1].

Milgram *et al.*'s seminal work on defining the Reality-Virtuality Continuum focuses mainly on how the choice of visual display technique influences our perception of the resulting mixed reality [11]. We propose a framework for classifying AR technologies for *all* the senses based on the location of where the real and CG elements are mixed. This approach complements Milgram's work by expanding the space within which multimodal display researchers can explore possible solutions.

Visual AR is typically implemented using one of three approaches, with the resulting images shown within a head-mounted display (HMD), or on another surface, such as the inside of a car windshield or cockpit. Video-see-through AR mixes a video stream captured from a camera mounted on the front of a fully-opaque HMD with CG images [1]. Optical-see-through AR uses an HMD with transparent or semi-transparent glass to overlay computer-generated images onto the user's view of the real world [3]. Here the mixing takes place in the visual field. The third method employs a head-mounted projection display (HMPD), which uses half-silvered mirrors to project CG images generated from the eye-point of the user into the environment [7]. Retro-reflective material is applied to objects that should reflect the projected light, thus rendering them "invisible" or transparent. Alternative approaches use projectors mounted in the real environment to insert annotations into the real environment [2].

In all of these examples, RW visual stimuli are mixed with CG imagery. In order to attain a truly merged experience, the two stimuli should undergo similar transformations, so that, for example, a virtual character receives the same lighting effects (light position and intensity) as objects in the real world. In fact, this applies to all sensory modalities; the voice of a virtual character should also be influenced by environmental objects, such as occluders or reflectors. Next, we dissect the paths that RW and CG stimuli travel from a source to a user, and describe the characteristics of AR stimuli that result when the mixing occurs at several interesting points along the paths.

2 Stimulus Pathways

We can look at the path that stimuli travel from their source to the user, determine points along these paths where the RW and CG elements mix, and analyze the characteristics of mixing at these locations. As the generation of RW stimuli is not under the control of the system, we assume they are created using some unknown mechanisms, and hence we only deal with the manifestations these stimuli can take as they approach the user. The generation of CG stimuli, on the other hand, is by definition under our control, and so the method of delivery (*i.e.*, display devices) often has a significant influence on the user experience. HMDs, for example, obscure some portion of the user's view of the real world, whereas projector-based displays do not [2].

1.1 Real-World Pathways

There are two general paths that a RW stimulus can take on its way to the user, one direct and one mediated. In the direct case (Figure 1a), a RW stimulus interacts with the surrounding environment (*e.g.*, is partially occluded by objects or structures) on its way to the appropriate sensory subsystem (*e.g.*, eyes, ears, skin), where it is translated into nerve impulses and transmitted to the brain. By inserting the desired CG elements into this path at different points (numbers 1-4 in Figure 1a), the resulting AR stimulus continues through the remaining stages on its way to the brain.

In the mediated case (Figure 1b), the RW stimulus again travels through the environment, but instead of being sensed by the user, it is captured by an appropriate sensing device (*e.g.*, camera, microphone, pressure sensor). The stimulus then undergoes optional post-processing, is mixed with CG elements, and is finally displayed to the user through appropriate display hardware at one of the points shown in Figure 1a.

VRST 2007, Newport Beach, California, November 5–7, 2007.
© 2007 ACM 978-1-59593-863-3/07/0011 $5.00

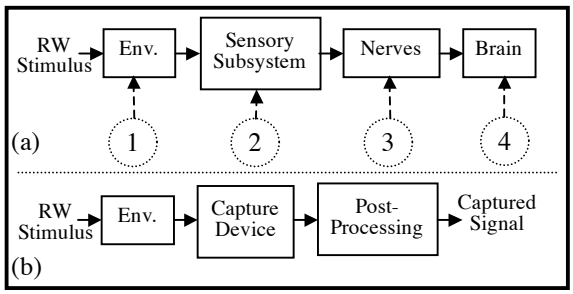

Figure 1: Pathways for real-world stimuli: (a) direct stimulus from the environment, (b) captured environmental stimuli.

1.2 Mixing Points

We have attempted to classify some AR display techniques for the various senses, based on the mixing points shown in Figure 1a. This classification is shown in Table 1, and includes examples for the Environmental mixing point (1) and the Sensory Subsystem mixing point (2). We are unaware of technologies for "displaying" sensory stimuli directly to the sensory nerves (optic, auditory, *etc.*) or directly to the brain.

Table 1: Examples of display technologies for various sensory modalities for the Environmental and Sensory mixing points.

Sense	**Environmental Display**	**Sensory Display**
Visual	Projection screens (far)	Retinal display
	Hand-held LCD (mid)	
	HMD (near)	
Auditory	Speakers (far/mid)	Bone conduction
	Headphones (near)	
Haptic	Subwoofer/floor/seat vibrators (mid)	Force reflector
	Fans (mid)	Pin arrays
	Heat lamp/Space heater (near)	Heating pad
Olfactory	Scent emitter (far/mid)	Mask-based
	Air canon (near)	
	Under-nose display (near)	
Gustatory	Edible displays (mid)	Tongue patch
	Taste tubes (near)	

The Environmental mixing point is further subdivided by the proximity of the display to the sensory subsystem (*e.g.*, visual, auditory). For example, an optical-see-through display where CG elements are projected onto the inside of a vehicle's window in front of the driver should be differentiated from a hand-held optical-see-through display, or a helmet-mounted optical-see-through display. This gradation is important because the closer CG elements are inserted to the sensory subsystem, the farther they are from the RW elements they need to be merged with, incurring an increase in complexity of the processing necessary to properly match the two.

To further clarify, the *earlier* the mixing point in the pathway, the *less* processing the CG element will require, as it can take advantage of the well-developed human sensory capabilities already utilized by the RW stimuli. For example, if a speaker array is used to insert CG audio into an environment (point 1), it will be altered by physical objects present in the environment, similar to the RW sound, on its way through the environment. If the audio is inserted at point 2, however, then 3D pre-processing of the CG audio, possibly using a model of the environment to compute proper occlusion, is required if the audio is to be perceptually merged at the user.

Determining the complexity of the delivery of captured stimuli is more problematic. The amount of processing necessary depends somewhat on how many stages there are between where the CG signal was acquired, and where it will be displayed. For example, microphones at the ears will capture a signal that will be very easy to play back to a user through headphones, but will require more complex computation to be played back on a set of speakers in a room.

3 AR Technologies for Mixing and Display

Looking at existing and envisioned techniques for mixing RW and CG stimuli across multiple modalities will help give a more-concrete view of the mixing-location continuum. Table 2 contains representative techniques for each sensory modality. Because the visual techniques were presented earlier, we focus now on describing the remaining modalities.

Table 2: Existing and envisioned technologies for AR across multiple sensory modalities.

		Location of Mixing		
		Environment	Sensory Subsystem	Computer
Sense	Visual	Optical-See-Through [3] HMPD [7], Projectors [2]	Retinal Display [14]	Video-See-Through [1]
	Audio	Speakers	Acoustic-Hear-Through	Microphone-Hear-Through
	Haptic	3D Printer	Actuated Stylus, Heating Pad, Vibrotactile Suit [9]	Exoskeleton
	Olfactory	Odor Emitter, Air Canon [16], Wearable [15, 13]		Mask-Based Display [6]
	Gustatory	Food-Mixing Device [5], Edible Bits [10]	Taste-Tube in Mouth	Feeding Tube in Mouth [8], Tongue patch

In the audio domain, CG sound can be displayed using speakers placed within the real environment, allowing both real and CG sounds to reach the user. Alternatively, CG sound can be displayed through headphones and, using two omni-directional microphones mounted on the outside of the headphones (Figure 2), environmental sound can be captured and mixed with the CG sound. We call this audio AR technique "Microphone-Hear-Through" AR (or simply *Mic-Through* AR). A third technique for audio AR can be thought of as "Acoustic-Hear-Through AR" (*Hear-Through* AR), which delivers CG sound through bone conduction [4], and RW sound through the unoccluded ear canals.

Figure 2: Microphones mounted on ear-bud headphones for capturing environmental audio.

Commercial bone-conducting devices have recently began to emerge onto the market. Figure 3 shows the AudioBone produced by Goldendance Co., Ltd., Japan. With this device, vibrational actuators are positioned on the zygomatic (cheek) bones in front of the ear. The unit has a normal output of 30mW, a maximum of 70mW, a normal impedance of 8 ohms, a sound-pressure sensitivity of 80 dB/mW (dB 1.0 dyne), and a standard operating frequency of 50Hz-4kHz. The total weight of the unit is 60g. The headband wraps around the back of the head, and the ear loops rest on the tops of the pinnae. Another innovative use of bone conduction is for listening to music while swimming (http://www.finisinc.com/products-swimp3v2.shtml), because bone conduction does not require sound to pass through air.

Figure 3: The AudioBone bone-conducting headset from Goldendance Co., Ltd.

The haptic sense is one of the most complex, as it encompasses both kinesthetic and cutaneous cues, as well as components of pain, temperature, and proprioception. In terms of force displays, an exoskeleton used for tele-operation could mix forces captured at the remote site with CG forces to produce resulting forces and torques displayed to the user. Alternatively, a stylus on the end of an actuated linkage, such as a PHANToM device, would combine the tactile sensations of grasping the stylus with the CG forces and torques generated when contacting CG objects. Thinking more abstractly, we can imagine a device that can produce and place physical objects into the real environment on demand, such as a 3D printer and a robot arm. This would mix these CG objects with other objects in the environment, allowing us to physically touch them. In terms of temperature, a computer-controlled fan could be used to blow hot or cold air at the user, in order to introduce information into the environment, such as proximity to a target location, or potential for trouble. A computer-controlled heating pad or peltier device could be used in the same manner, but would be more proximal, so as not to disturb others, and to give cues more rapidly.

In the olfactory domain, a computer-controlled scent emitter placed within a real environment would allow a user to receive a mix of scents from the emitter and the physical environment. Alternatively, an oxygen-mask-type olfactory display [6] could be used to mix captured smell from the environment with CG scents, though the synthesis of scents remains an elusive task. Projection [16] or wearable [15] olfactory displays allow environmental and CG scents to mix at the nasal passage [13].

We can even envision the same classification technique being applied to displaying to the gustatory (taste) sense. Similar to a 3D printer, a food dispenser that places food in the environment would allow the placement of food to be computer controlled. While this is clearly a multimodal interface (vision to see the food, haptics to grasp and chew it [8], olfactory to smell it, and gustatory to taste it), this closely mimics the real eating experience. Alternatively, a food mixer would allow us to alter the taste of existing food within the environment. A feeding tube in the mouth could be used to deliver actual food, or flavored liquid, to the mouth [8]. Such a tube could either be used in place of actual food, or be used in conjunction with real food (*i.e.*, a "taste-tube") in the mouth. We can envision a type of patch worn on the tongue to release various tastes from a taste cartridge. A straw-like user interface has been proposed [5] that combines tactile, olfactory, and gustatory sensory stimulation to present an augmented drinking experience, mostly under computer control. Edible user interfaces have been proposed that display food on the surface of a computer screen, and allow users to "taste" the current state of the system [10]. Because of its reliance on olfaction, the intrusive nature of potential delivery systems, and the complex nature of the eating process, a general approach to stimulating the sense of taste remains a difficult problem.

4 Discussion

Though some of the "displays" in the preceding discussion may seem futuristic (and cumbersome!), the section was meant to give an illustration of the applicability of the mixing-point continuum. Some observations about the continuum further help in the discussion.

First, in general, RW stimuli offer high fidelity at relatively low computational cost, but provide less control over what the user actually experiences. CG stimuli, on the other hand, offer almost total control, but with a direct relationship between fidelity (or quality) and computational cost: as fidelity increases, so does cost. Stimuli captured from the real world, such as photographic textures, sampled sound, or motion-capture data, provide realistic results due to the fact that they are taken from reality. However these data provide only limited processing options. Synthetic textures and sound, or motion created by an animator, allow the resulting stimuli to be precisely specified, but mirroring reality can be challenging.

Another characteristic is that the later the mixing point, the more personal are the cues. "Public" displays, such as projection screens, speakers, or air-canon olfactory displays, provide stimuli to anyone within range, while HMDs, bone-conduction headphones, and mask-type olfactory displays feed their stimuli only to the immediate user. The appropriateness of a particular display will vary based on the application, the usage environment, and/or cost.

Because life is a multimodal experience, addressing individual sensory stimuli in isolation will probably not be the most efficient means of providing a seamless AR experience. It is important for systems to address how mixing location and technology choices made for one sensory modality constrain the options for the remaining modalities. A difference in stimulus quality for one modality may either enhance or reduce the perceived quality of a stimulus delivered to another modality. For example, adding visual deformation to the CG representation of a rigid physical surface explored with the user's real hands could lead the user to perceive an elastic surface, thereby reducing the need to provide a computationally expensive, deformable haptic surface. If, however, the cues are in large conflict with each other, the overall experience may be diminished; sensory dominance should therefore be taken into consideration.

On the other hand, we *can* consider each signal in isolation as long as the resulting information the brain receives is in concert. We believe achieving harmony will be more difficult for multiple signals of the same modality (*e.g.*, audio) following different pathways, than for signals for different senses, as accurately matching the virtual environment to the real environment requires great care. For example, for environmental audio delivered from speakers to be perceived as existing in the same space as voice from a virtual character delivered via a bone-conduction device, the voice signal should be processed using the same "environmental effects" the speaker audio undergoes as it travels from the speakers to the listener. We believe that users will be more sensitive to differences of this type as opposed to cross-modal differences, as shortcomings in one will be masked by the other. This is clearly an interesting area of future study.

5 Conclusion

One of our longer-term goals is the creation and delivery of authentic multimodal AR stimuli. That is, we aim to produce a stimulus that combines computer-generated or mediated cues with environmental cues in such a way that users will successfully merge RW and CG elements. While efforts to provide stimuli directly to the nerves or brain are scarce, found mostly in works of popular science fiction, they might hold the key to crossing the Uncanny Valley [12]. Until we can successfully tap into the most proximal mixing points on the path to the brain, it is hoped that the considerations raised here will allow AR system designers to select the most appropriate techniques. Finally, we hope our framework will suggest new techniques that can be empirically evaluated with regard to how well users perceive the AR stimuli as providing a seamless experience.

Acknowledgements

This research was supported in part by the National Institute of Information and Communications Technology of Japan.

References

[1] Billinghurst, M., Kato, H. Collaborative Augmented Reality, *Comm. of the ACM*, 45(7), 2002, 64-70.

[2] Ehnes, J., Hirota, K., Hirose, M. Projected Augmentation - Augmented Reality using Rotatable Video Projectors, *Proc. of the 3rd IEEE and ACM Int'l Symp. on Mixed and Augmented Reality (ISMAR '04)*, 26-35.

[3] Feiner, S., MacIntyre, B., Haupt, M., Solomon, E. Windows on the World: 2D Windows for 3D Augmented Reality, *Proc. of UIST '93*, 145-155.

[4] Fukumoto. M., Tonomura, Y. Whisper: A Wristwatch Style Wearable Handset, *Proc. of ACM CHI '99*, 112-119.

[5] Hashimoto, Y., Nagaya, N., Kojima, M., Miyajima, S., Ohtaki, J., Yamamoto, A., Mitani, T., Inami, M. Straw-like User Interface: Virtual Experience of the Sensation of Drinking Using a Straw, *Proc. of the 2006 ACM SIGCHI Int'l Conf. on Advances in Comp. Entertainment Tech (ACE '06)*, 2006.

[6] Hirose, M., Tanikawa, T., Ishida, K. A Study of Olfactory Display, *Proc. of the Virtual Reality Soc. of Japan 2nd Annual Conf. 1997* (in Japanese), 155-158.

[7] Inami, M., Kawakami, N., Sekiguchi, D., Yanagida, Y., Maeda, T., Tachi, S. Visuo-Haptic Display Using Head-Mounted Projector, *Proc. of IEEE Virtual Reality 2000*, 233-240.

[8] Iwata, H., Yano, H., Uemura, T., Moriya, T. Food Simulator: A Haptic Interface for Biting, *Proc. IEEE Virtual Reality 2004*, 51-57.

[9] Lindeman, R.W., Yanagida, Y., Noma, H., Hosaka, K. Wearable Vibrotactile Systems for Virtual Contact and Information Display, *Virtual Reality*, 9(2-3), 2006, 203-213.

[10] Maynes-Aminzade, D. Edible Bits: Seamless Interfaces Between People, Data, and Food, *ACM CHI 2005 Extended Abstracts*, 2207-2210.

[11] Milgram, P., Takemura, H., Utsumi, A., Kishino, F. Augmented Reality: A Class of Displays on the Reality-Virtuality Continuum, *SPIE Vol. 2351, Telemanipulator and Telepresence Technologies*, 1994, 282-292.

[12] Mori, M. The Uncanny Valley, *Energy*, 7(4), 1970, 33-35 (translated by K.F. MacDorman and T. Minato)

[13] Nakamoto T., Min, P.H.D. Improvement of Olfactory Display Using Solenoid Valves, *Proc. of IEEE Virtual Reality 2007*, 179-186.

[14] Pryor, H.L., Furness, T.A., Viirre, E. The Virtual Retinal Display: A New Display Technology Using Scanned Laser Light, *Proc. of Human Factors and Ergonomics Soc., 42nd Annual Meeting*, 1998, 1570-1574.

[15] Yamada, T., Yokoyama, S., Tanikawa, T., Hirota, K., Hirose, M. Wearable Olfactory Display: Using Odor in Outdoor Environment, *Proc. of IEEE Virtual Reality 2006*, 199-206.

[16] Yanagida, Y., Kawato, S., Noma, H., Tomono, A., Tetsutani, N. Projection-Based Olfactory Display with Nose Tracking, *Proc. of IEEE Virtual Reality 2004*, 43-50.