(Re-)Examination of Multimodal Augmented Reality

Nina Rosa
Utrecht University
Utrecht, The Netherlands
n.e.rosa@uu.nl

Peter Werkhoven
Utrecht University
Utrecht, The Netherlands
p.j.werkhoven@uu.nl

Wolfgang Hürst
Utrecht University
Utrecht, The Netherlands
huerst@uu.nl

ABSTRACT

The majority of augmented reality (AR) research has been concerned with visual perception, however the move towards multimodality is imminent. At the same time, there is no clear vision of what multimodal AR is. The purpose of this position paper is to consider possible ways of examining AR other than using popular notions and definitions, in order to come to an understanding of multimodal AR. Three concepts are introduced. First, the commonly used terms 'real' and 'virtual' are redefined in terms of stimuli rather than objects, and mediated stimuli are introduced as a new category. This addition facilitates the support of, for example, sensory substitution stimuli in AR. Next, the popular definitions of AR are examined, and a new analysis is presented. Specifically, it is theorized that the core of AR is not necessarily the combination of real and virtual, but the combination of a basis and an augmentation. Such an abstraction allows the re-introduction of augmented virtuality and, more generally, mixed reality. Lastly, a classification system for different forms of multimodal mixed reality that applies this basis-augmentation model is presented.

CCS Concepts

•Human-centered computing → Mixed / augmented reality;

Keywords

Augmented reality; multimodality; perception

1. INTRODUCTION

Augmented reality (AR) is commonly understood as a real environment that is augmented by virtual elements. Although there have been efforts in the past to concretely define AR and the elements it contains [10, 3], there are still cases of which the correct classification is debatable. One such case, for example, is whether digitally captured and displayed elements are real or virtual. Another such case is whether an application in which real occurs in one modality and the virtual in another should also be considered AR. Because of such debates we do not have a clear vision of what types of AR applications exist, or how fields have progressed since the introduction of AR, or what the future of AR research may hold.

The purpose of this position paper is *not* to throw away all known, recognized definitions and ideas about AR, but to examine them and discuss uncertainties that arise. In particular, it is argued that the elements of AR are not only real and virtual stimuli, but also mediated stimuli (Section 2). Moreover, the core of AR is not simply the combination of these stimuli, but more generally a physical basis that is augmented by non-physical elements (Section 3). Viewing AR in such way allows us to reconsider the notion of augmented virtuality (AV), and in turn the notion of mixed reality (MR). By considering previous works on classifying multimodal AR (Section 4), a classification system for multimodal MR is presented based on the basis-augmentation model (Section 5).

2. REAL, MEDIATED, VIRTUAL STIMULI

Although AR research has an immense corpus of literature, very little work has been dedicated to the exact differentiation between the terms real and virtual (which may in fact reflect the difficulty of this topic). Milgram and Kishino [10] realize the need for such a distinction, as many discussions occurred when examining various aspects of MR applications. They define real and virtual objects along three dimensions. Firstly, a real object exists objectively, whereas a virtual object exists in essence but not formally. Secondly, real objects can be viewed directly or indirectly, e.g. through a sensor-display mediator, whereas virtual objects can only be viewed indirectly. Lastly, a distinction is made between real and virtual images: real images have correct luminosity considering their surroundings, whereas virtual images have no luminosity or are transparent.

The authors state that even this definition is not sufficient. In particular, it is noted that it is strange to consider both a remotely viewed video scene and one's own directly viewed hand as real, but do not further elaborate on this discrepancy. Some authors nowadays even address the former as virtual. This discrepancy is not further elaborated, but a similar discussion can be found in two other works, namely those of Mann [9] and Müller [11]. Mann argues that a mediated view of the world (where mediation is not restricted to digitalization) is fundamentally different from a direct, undistorted view of the world. Therefore, any environment

MVAR'16, November 16 2016, Tokyo, Japan

DOI: http://dx.doi.org/10.1145/3001959.3001961

can not only be placed on a real-virtual continuum (in contrast to [10]), but also on a real-mediated continuum. Müller classifies information in procedural tasks in AR into five layers: the real world, the mediated world, virtual objects that are spatially referenced and of spatial nature, virtual objects that are spatially referenced but not of spatial nature, and virtual objects that do not have any connection to the physical world.

In essence there are two issues at hand here. Firstly, while AR is commonly described as consisting of real and virtual elements, in terms of perception and experience mediated elements are something different and should be considered as a separate type. Secondly, Milgram & Kishino and Müller (in the first three layers) both discuss real and virtual (and mediated) in terms of objects in AR, while Mann discusses the environments. Since today AR is not restricted to only augmentation of the real environment by virtual objects, but also allows for example modification of the real by virtual, a more appropriate element of discussion may be those elements that make up the perception of our environment, namely stimuli. Of course, this still allows the examination of complete objects, for example, in the visual sense, but also allows the examination of one's environment through other modalities.

So what is a mediated stimulus in AR? Translating Mann's description of mediated to AR only, it is a digitally modified stimulus. The trivial case is using a sensor-display combination for illusory transparency. That is, it is currently not possible to exactly reconstruct the physical, and there is always some form of inconsistency and therefore modification. Examples of non-trivial modification, on the other hand, are: looking at your arm where the skin color has been changed digitally, or hearing someone's 'roboticized' voice, or sensing tactile sensations slightly intensified (Bayart and Kheddar define this as enhanced haptic [5]). A more abstract example of modification would be sensory substitution, where a stimulus for a certain modality is presented as a stimulus for a different modality. A well-known example is using vibrotactile arrays on the skin, where a vibration motor is activated according to a light intensity measured by a camera, in order to 'see with the skin'. Examples of sensory substitution are widespread; see [4] for an overview. The concept of sensory substitution can be extended to stimuli that humans cannot perceive (haptic enhancing in [5]). Is it useful to point out that these forms of sensory substitution can also occur for virtual stimuli; an example can be found in *Luigi's Ghost Mansion* where the player senses when virtual ghosts are in close proximity through vibrations [1].

To limit the scope of 'modification' (i.e. the border between mediated and virtual), it is required that the modification is not random, and that there is a clear relationship between the modified and non-modified form. To illustrate, a point-cloud representation of a real person is considered mediated, whereas a 3D model of a generic human that is not derived from the real person in question is considered virtual, and of course the direct view of the person is real.

Putting together the above concepts into definitions, we have the following:

Real the stimulus originates from the physical environment, and is perceived without any form of mediation, by its intended modality.

Mediated the stimulus originates from the physical environment, and is perceived through a digital sensor-display mediator, either by the intended modality (for which it may have been altered) or other modality(ies).

Virtual the stimulus originates from a computer-generated model, and is perceived through a digital display, either by the intended modality or other modality(ies).

There are a few benefits to the differentiation between real, mediated and virtual. Firstly, it allows all applications for which the virtual means either 'digitally presented information' and/or 'computer-generated information' to still be considered AR (this is further elaborated in Section 3). Secondly, sensory substitution stimuli, an important concept in multimodality which were formerly not considered real or virtual (with the exception being [5]), can now be classified as an element in AR. More generally, any real stimulus that is digitally modified in some way can now be classified within AR. Lastly, it is likely that there is a difference regarding experience between the three types. For real/mediated vs. virtual, this is clear: the user knows that the former originates from the physical world, so it is experienced as 'more real' than the virtual. It is not strange to suggest that such a difference may also occur between real and mediated. For example, in a scenario where a user's arm is visually mediated and the other is visually real, will the user react more strongly to a threat to the real arm than to a threat to the mediated arm?

An important remark should be made here. It can already be the case that user's cannot decide whether something is mediated or virtual, due to the extent of modification. What is more, as technology progresses and becomes more integrated into our daily lives, it may be that at one point humans can no longer tell the difference between what is and what is not physical in the perceptual sense, meaning the experiences for real, mediated, and virtual could all be identical. Indeed, one could even argue that this all comes down to prior knowledge of one's environment. This utopian situation however is not expected anywhere in the near future, thus for the time being, we assume this difference can be detected.

3. BASIS & AUGMENTATION

In the previous section, the elements of AR were respecified as opposed to the popular definition [10]. Following this discussion, the popular definition of AR is evaluated. The purpose is not to invalidate previous works, but to incorporate the notion of real, mediated and virtual stimuli.

There are two influential works regarding the definition of AR: that of Milgram and Kishino [10] and that of Azuma [3]. Milgram and Kishino are widely known for their Reality-Virtuality Continuum, and define AR as all cases in which the display of an otherwise real environment is augmented by means of virtual objects, and is a subset of MR on the continuum. Azuma states that, indeed, a property for an AR system is that it combines real and virtual objects in a real environment. Furthermore he specifies that it must run interactively and in real-time, and it must register/align real and virtual objects with each other.

Both definitions agree in two aspects: the basis of AR must be a 'real' environment, and the augmentation is 'virtual'. According to the new stimuli-framework, there are two options for the basis, namely the basis is real or it

is mediated. This distinction can already be found in the most common forms of visual AR technology: see-through head-mounted displays (HMDs) provide a real environment as a basis, whereas video-based displays provide a mediated environment as a basis. The stimuli-framework then also allows augmentation by both mediated and virtual stimuli. To evaluate all options, it is useful to consider the possible combinations of basis-augmentation within the context of a single application scenario. For this purpose the *Holoportation* technology by Microsoft is used as an example. This 3D capture technology allows precise, real-time reconstruction of humans, and in combination with AR displays can 'allow interaction' with remote participants in 3D [2]. The demonstration shows a father wearing a see-through AR head-mounted display (HMD) that can see a digital representation of his daughter (that is located in a different room) as if in the same room. Now, for each basis-augmentation combination, the following variations of this scenario hold:

- Real-Mediated - see-through HMD AR where the interaction is with a real remote participant
- Real-Virtual - see-through HMD AR where the interaction is with a virtual agent (rather than a real remote participant)
- Mediated-Mediated - video-based AR where the interaction is with a real remote participant
- Mediated-Virtual - video-based AR where the interaction is with a virtual agent
- Real-Mediated-Virtual - see-through HMD AR where the interaction is with a real remote participant and a virtual agent.

In the last example, the mediated stimuli are part of the augmentation but generally they can be part of the basis, augmentation, or both. Although the combinations are sketched for one single application scenario, it should be clear that these variations hold for (almost) every other (perceptual) AR application.

What has been shown here is that the previous definitions of AR still hold to some degree with the stimuli-framework; there is a physical basis, and a non-physical augmentation. The difference is that this basis need not be strictly real, and the augmentation need not be strictly virtual. In other words, the focus of AR has been shifted to a more generalized concept, namely basis-augmentation rather than real-virtual. In this way, AR is more clearly defined within MR, and AV can be reintroduced [10]. A requirement is that the basis and augmentation are related to the context of the application; simply adding unrelated virtual media does not validate the use of the term AR. In addition, complying with Azuma's other properties, we require that the AR system is interactive at real-time, and that real and virtual stimuli (rather than objects per se) are correctly registered with each other. By the latter is meant that, for example, the real and virtual visuals must be correctly registered with each other, but also that the location of the virtual visuals matches the location of the related virtual haptics.

So far only the definitions of unimodal visual AR have been examined. In the next sections, the concept of multimodal AR is introduced, together with a new classification system for multimodal AR (or, more generally, MR) based on the presented basis-augmentation model.

4. MULTIMODAL AUGMENTED REALITY

Proposed concepts in visual AR literature are often said to be similarly applicable in multimodal situations. However, such conjectures are hardly ever verified which can lead to contradicting views. For example, imagine a navigation system where the visual directions are overlaid onto the real visuals. Many would consider this system AR. Now, imagine the following alteration: the directions are presented as vibrations around the waist, similar to what is described in [13]. In the context of navigation, the visuals are real and the directions are virtual. Although the alteration is slight, very few people would consider this version AR. Milgram and Kishino consider this case so different that they suggest using a new term altogether, namely 'hybrid reality' [10]. This contrast is caused by the multimodal factor.

A small group of works have been dedicated to the closer examination of multimodal AR, and in particular, the larger number of variations it allows compared to unimodal AR. Kalawsky et al. propose a framework based on a functional decomposition of AR [7]. The purpose is to allow complete sensory description of the user's capabilities to compare systems. Although the taxonomy can be used to describe a specific multimodal AR application, it is in fact not specific for multimodality nor for AR. It is a decomposition based on any generic human-computer interface, and there is no mention of 'combining real and virtual' or 'augmentation' for that matter. Lindeman and Noma provide a classification system for AR based on where the real-virtual mixing occurs, which can be used for each sense individually [8]. They give examples of technologies for various mixing locations and describe important implications of mixing location, such as the fact that mixing location and technology choices made for one modality can constrain the options for the remaining modalities. However, the classification system is restricted to those cases where the virtual must be as realistic as possible and does not consider, for example, more abstract forms of sensory substitution. Jeon and Choi define a visuo-haptic MR taxonomy [6], which consists of two orthogonal reality-virtuality continua (from [10]), one for visuals and one for haptics. Unfortunately, it extends poorly to all modalities, since the complexity grows exponentially with each added modality.

What these works so far do not illustrate is that multimodal AR can indeed mean that the basis and augmentation are within multiple modalities, but also that they are spread across modalities. This lack of 'spread multimodality' in the literature may reflect natural expectations humans have for their perception by the senses in the real world, such as perceiving in high quality and congruently across all relevant modalities. Previous visual AR research has been more concerned with correct implementation of the first aspect, and when the goal of an AR application is to recreate a real scenario, then we must also concern ourselves with the second characteristic. However, this goal is not necessarily generalizable to all AR applications.

Because the concept of AR has now been modeled around modalities, a limitation occurs. Namely, a small group of scenarios of which the real and virtual are not at a perceptual level (that is, they are not stimuli) would no longer be regarded as AR, although there are currently considered as such. Examples include location-based AR and natural language-based scenarios, since these both require the user to process at a cognitive level. We do not claim that these

Table 1: The classification of different types of multimodal mixed reality, illustrated by augmented reality examples using vision, audio and/or haptics.

MR type	Vision		Audio		Haptics		AR Example
	Basis	Augm.	Basis	Augm.	Basis	Augm.	
Intramodal	✓	✓	✓	✓			Holoportation [2]
Intermodal	✓	✓			✓		Soft AR [12]
	✓	✓				✓	Luigi's Ghost Mansion [1] for AR
	✓	✓	✓			✓	Soccer Scenery
Crossmodal	✓					✓	Visual Navigation with Tactile Belt [13]

examples are therefore not AR, but simply that they do not fit in a model working towards multimodal AR.

5. INTER-, INTRA-, CROSSMODALITY

As stated earlier, the basis-augmentation model from Section 3 describes both AR and AV, the two subsets of MR. Therefore, a classification system based on this model can be applied to MR in general. In this system, the different categories depend on the degree of spreadness of basis and augmentation across modalities. The classification is therefore not concerned with the quality of the stimuli, the number of relevant utilized modalities, or the quality of blending of the basis and augmentation. All that is required is that the basis and augmentation are linked by the purpose of the application (as before). Table 1 illustrates the classification and gives examples of AR applications for each category, which will each be further elaborated in the following. These examples only regard vision, audio, and/or haptics, but the classification works for all modalities, including those outside of the traditional five. Only AR examples are given for simplicity, and these were chosen based on the different experiences they provide.

To start, there are at least two cases: the basis and augmentation are within modalities (e.g. basis and augmentation in both vision and audition) or they are not (e.g. visual basis and auditory augmentation). These two cases are called intramodal and crossmodal MR, respectively. *Intramodality* is the generalization of unimodal MR to multimodal MR, where for each modality related to the application there is 'full MR'. Generally, intramodality is useful for scenarios where the purpose is to have the final augmented scene as close to a realistic one as possible. Returning to the Holoportation example, the purpose of AR in this scenario is to create the sensation that the daughter is actually present alongside the father, and therefore it is important to include both basis and augmentation in those modalities that matter. In the demonstration video, the father sees and hears this mediated daughter, which creates the impression she is there. *Crossmodality* requires that the modalities of the basis and augmentation are mutually exclusive, that is, there is no overlap in modalities. An example of this is the visuotactile navigation scenario described earlier. Of course, in this case the user is still capable of perceiving other real stimuli, such as real tactile sensations, however these stimuli are not crucial to the application, so they are not considered as a basis modality. In this case, the purpose of crossmodality is to ensure high performance for certain tasks, without overly increasing the stimuli of a single modality. For example, a known problem in visual AR is the clutter caused by the provided amount of virtual stimuli which may increase the stress-levels or overall comfort of the user. Moving some information to other modalities is a possible solution to this.

The last type of MR is the middle ground between 'not any' and 'complete' spreadness across modalities. *Intermodality* indicates that there is intramodality, and either a basis in one or more different modalities, an augmentation in one or more different modalities, or crossmodality. An example of the first subtype is SoftAR, where a user can see and touch a real object, and virtual indent marks are projected onto the object such that the user experiences it as softer than it really is [12]. An example of the second type would be an AR variation of Luigi's Ghost Mansion, where a player has to locate the virtual ghosts, of which the presence is felt through the tactile sense, and they appear and are defeated when the player looks straight at them. Lastly, an example of the third type would be a real game of soccer but the surroundings of the field are augmented to represent a virtual stadium where the fans are chanting.

To summarize, the classification characterizes the following types of MR, from least to most spreadness across modalities:

Intramodality all bases and augmentations are within the same modalities

Intermodality intramodality in at least one modality and:

- **type 1** a basis in one or multiple different modalities
- **type 2** an augmentation in one or multiple different modalities
- **type 3** crossmodality in other modalities

Crossmodality the modalities of the bases and augmentations are mutually exclusive.

This classification is useful to gain insight on different areas of research being conducted in multimodal AR. It is generalizable to all modalities, and can be restricted when necessary to a specific group of modalities as done in Table 1 and in [6] (further elaborated below). Another benefit is that it can be linked to application purposes and used as a tool to understand the minimal requirements of the system. For example, when a MR scene is required that mimics the real world as close as possible, intramodality is likely desired. On the other hand, when the goal is to increase performance of a certain task, crossmodality offers valuable implementation options. Intermodality may be used when a combination of these two goals is necessary. We emphasize that these are rough generalizations, and recognize that exceptions exist.

The goal of this classification is not necessarily to replace all earlier mentioned classifications but to complement them creating a more overarching view of what multimodality in

MR has to offer. For example, the taxonomy described by Jeon and Choi [6] is very similar in nature, but a few key differences are noticeable. Firstly, the cases that they describe as MR are actually only cases where there is intra- or intermodality (types 1 and 2). The currently proposed classification disagrees and would also consider 'rV-vH' and 'vV-rH' as MR, specifically, crossmodal MR. Secondly, the third subtype of intermodality does not exist in their taxonomy, because it only considers two modalities at a time.

6. CONCLUSION

The goal of this position paper was to introduce new ways to think about AR. In particular, the notion of real and virtual were respecified in terms of stimuli to emphasize physicality and perception, and a new form was introduced, namely mediated stimuli. These stimuli originate from the physical world but are captured and displayed to the user in the intended modality or another. It was shown that this distinction allows the incorporation of sensory substitution as a type of stimulus and in fact permits agreement on the variety of applications that are currently considered AR. This is done by regarding AR as a combination of a real and/or mediated basis, and a mediated and/or virtual augmentation. This basis-augmentation showed to be appropriate terminology when examined with respect to the larger concept of MR. Lastly, it was reasoned that multimodal AR can take on different forms with respect to spreadness across modalities: the basis and augmentation can be within multiple modalities, but also spread across modalities. Following this rationale, a classification system for multimodal MR was presented. We acknowledge that there are still many issues regarding the presented classification, however it remains a useful tool for the upcoming research field of multimodal MR.

7. REFERENCES

[1] Luigi's Ghost Mansion, Nintendo Land, Nintendo Wii U. Video Game, 2012.

[2] Holoportation by Microsoft Research. https://www.microsoft.com/en-us/research/project/holoportation-3/, 2016. Accessed: 2016-07-20.

[3] R. Azuma, Y. Baillot, R. Behringer, S. Feiner, S. Julier, and B. MacIntyre. Recent advances in augmented reality. *IEEE computer graphics and applications*, 21(6):34–47, 2001.

[4] P. Bach-y Rita and S. W. Kercel. Sensory substitution and the human–machine interface. *Trends in cognitive sciences*, 7(12):541–546, 2003.

[5] B. Bayart and A. Kheddar. Haptic augmented reality taxonomy: haptic enhancing and enhanced haptics. In *Proceedings of EuroHaptics*, pages 641–644. Citeseer, 2006.

[6] S. Jeon and S. Choi. Haptic augmented reality: Taxonomy and an example of stiffness modulation. *Presence: Teleoperators and Virtual Environments*, 18(5):387–408, 2009.

[7] R. Kalawsky, A. Stedmon, K. Hill, and C. Cook. A taxonomy of technology: Defining augmented reality. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 44, pages 507–510. SAGE Publications, 2000.

[8] R. W. Lindeman and H. Noma. A classification scheme for multi-sensory augmented reality. In *Proceedings of the 2007 ACM symposium on Virtual reality software and technology*, pages 175–178. ACM, 2007.

[9] S. Mann. Mediated reality with implementations for everyday life. *Presence Connect*, 2002. Posted: August 6, 2002.

[10] P. Milgram and F. Kishino. A taxonomy of mixed reality visual displays. *IEICE TRANSACTIONS on Information and Systems*, 77(12):1321–1329, 1994.

[11] T. Müller. Towards a framework for information presentation in augmented reality for the support of procedural tasks. In *International Conference on Augmented and Virtual Reality*, pages 490–497. Springer, 2015.

[12] P. Punpongsanon, D. Iwai, and K. Sato. SoftAR: Visually manipulating haptic softness perception in spatial augmented reality. *IEEE transactions on visualization and computer graphics*, 21(11):1279–1288, 2015.

[13] J. B. Van Erp, H. A. Van Veen, C. Jansen, and T. Dobbins. Waypoint navigation with a vibrotactile waist belt. *ACM Transactions on Applied Perception (TAP)*, 2(2):106–117, 2005.