IEICE Transactions on Information Systems, Vol E77-D, No.12 December 1994.

A TAXONOMY OF MIXED REALITY VISUAL DISPLAYS

Paul Milgram º
Department of Industrial Engineering
University of Toronto
Toronto, Ontario, Canada M5S 1A4

+

Fumio Kishino ºº
ATR Communication Systems Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun
Kyoto 619-02, Japan

Manuscript received July 8, 1994.
Manuscript revised August 25, 1994.

º *The author is with the Department of Industrial Engineering, University of Toronto, Toronto, Ontario, Canada M5S 1A4.*

- **Paul Milgram** received the B.A.Sc. degree from the University of Toronto in 1970, the M.S.E.E. degree from the Technion (Israel) in 1973 and the Ph.D. degree from the University of Toronto in 1980. From 1980 to 1982 he was a ZWO Visiting Scientist and a NATO Postdoctoral in the Netherlands, researching automobile driving behaviour. From 1982 to 1984 he was a Senior Research Engineer in Human Engineering at the National Aerospace Laboratory (NLR) in Amsterdam, where his work involved the modelling of aircraft flight crew activity, advanced display concepts and control loops with human operators in space teleoperation. Since 1986 he has worked at the Industrial Engineering Department of the University of Toronto, where he is currently an Associate Professor and Coordinator of the Human Factors Engineering group. He is also cross appointed to the Department of Psychology. In 1993-94 he was an invited researcher at the ATR Communication Systems Research Laboratories, in Kyoto, Japan. His research interests include display and control issues in telerobotics and virtual environments, stereoscopic video and computer graphics, cognitive engineering, and human factors issues in medicine. He is also President of Translucent Technologies, a company which produces "Plato" liquid crystal visual occlusion spectacles (of which he is the inventor), for visual and psychomotor research.

ºº *The author is with ATR Communication Systems Research Laboratories, Kyoto-fu, 619-02 Japan.*

- **Fumio Kishino** is a head of Artificial Intelligence Department, ATR Communication Systems Research Laboratories. He received the B.E. and M.E. degrees from Nagoya Institute of

Technology, Nagoya, Japan, in 1969 and 1971, respectively. In 1971, he joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Corporation, where he was involved in work on research and development of image processing and visual communication systems. In mid-1989, he joined ATR Communication Systems Research Laboratories. His research interests include 3D visual communication and image processing. He is a member of IEEE and ITEJ.

Summary

This paper focuses on Mixed Reality (MR) visual displays, a particular subset of Virtual Reality (VR) related technologies that involve the merging of real and virtual worlds somewhere along the "virtuality continuum" which connects completely real environments to completely virtual ones. Probably the best known of these is Augmented Reality (AR), which refers to all cases in which the display of an otherwise real environment is augmented by means of virtual (computer graphic) objects. The converse case on the virtuality continuum is therefore Augmented Virtuality (AV). Six classes of hybrid MR display environments are identified. However, an attempt to distinguish these classes on the basis of whether they are primarily video or computer graphics based, whether the real world is viewed directly or via some electronic display medium, whether the viewer is intended to feel part of the world or on the outside looking in, and whether or not the scale of the display is intended to map orthoscopically onto the real world leads to quite different groupings among the six identified classes, thereby demonstrating the need for an efficient taxonomy, or classification framework, according to which essential differences can be identified. The 'obvious' distinction between the terms "real" and "virtual" is shown to have a number of different aspects, depending on whether one is dealing with real or virtual objects, real or virtual images, and direct or non-direct viewing of these. An (approximately) three dimensional taxonomy is proposed, comprising the following dimensions: Extent of World Knowledge ("how much do we know about the world being displayed?"), Reproduction Fidelity ("how 'realistically' are we able to display it?"), and Extent of Presence Metaphor ("what is the extent of the illusion that the observer is present within that world?").
key words: *virtual reality (VR), augmented reality (AR), mixed reality (MR)*

---

1. Introduction -- Mixed Reality

The next generation telecommunication environment is envisaged to be one which will provide an "ideal virtual space with [sufficient] reality essential for communication"º . Our objective in this paper is to examine this concept, of having both "virtual space" on the one hand and "reality" on the other available within the same visual display environment.

The conventionally held view of a *Virtual Reality* (VR) environment is one in which the participant-observer is totally immersed in, and able to interact with, a completely synthetic world. Such a world may mimic the properties of some real-world environments, either existing or fictional; however, it can also exceed the bounds of physical reality by creating a world in which the physical laws ordinarily governing space, time, mechanics, material properties, etc. no longer hold. What may be overlooked in this view, however, is that the VR label is also frequently used in association with a variety of other environments, to which total immersion and complete synthesis do not necessarily pertain, but which fall somewhere along a *virtuality continuum.* In this paper we focus on a particular subclass of VR related technologies that involve the merging of real and virtual worlds, which we refer to generically as *Mixed Reality (MR)*. Our objective is to formulate a taxonomy of the various ways in which the "virtual" and "real" aspects of MR environments can be realised. The perceived need to do this arises out of our own experiences with this class of environments, with respect to which parallel problems of inexact terminologies and unclear conceptual boundaries appear to exist

among researchers n the field.

The concept of a "virtuality continuum" relates to the mixture of classes of objects presented in any particular display situation, as illustrated in Figure 1, where *real environments*, are shown at one end of the continuum, and *virtual environments*, at the opposite extremum. The former case, at the left, defines environments consisting solely of real objects (defined below), and includes for example what is observed via a conventional video display of a real-world scene. An additional example includes direct viewing of the same real scene, but not via any particular electronic display system. The latter case, at the right, defines environments consisting solely of virtual objects (defined below), an example of which would be a conventional computer graphic simulation. As indicated in the figure, the most straightforward way to view a Mixed Reality environment, therefore, is one in which real world and virtual world objects are presented together within a single display, that is, anywhere between the extrema of the virtuality continuum.

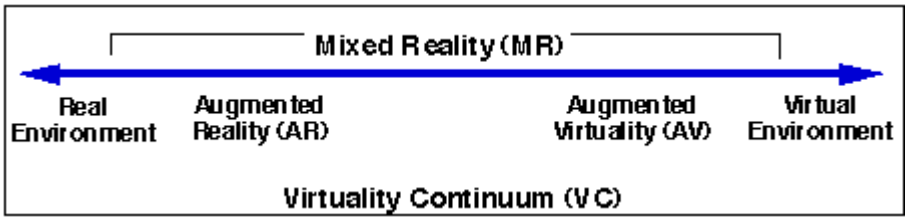

Figure 1: Simplified representation of a "virtuality continuum".

Although the term "Mixed Reality" is not (yet) well known, several classes of existing hybrid display environments can be found, which could reasonably be considered to constitute MR interfaces according to our definition:

- 1. Monitor based (non-immersive) video displays – i.e. "window-on-the-world" (WoW) displays – upon which computer generated images are electronically or digitally overlaid (e.g. Metzger, 1993; Milgram et al, 1991; Rosenberg, 1993; Tani et al, 1992). Although the technology for accomplishing such combinations has been around for some time, most notably by means of chroma-keying, practical considerations compel us to be interested particularly in systems in which this is done *stereoscopically* (e.g. Drascic et al, 1993; Lion et al, 1993).

- 2. Video displays as in Class 1, but using immersive head-mounted displays (HMD's), rather than WoW monitors.

- 3. HMD's equipped with a see-through capability, with which computer generated graphics can be optically superimposed, using half-silvered mirrors, onto directly viewed real-world scenes (e.g. Bajura et al, 1992; Caudell & Mizell, 1992; Ellis & Bucher, 1992; Feiner et al, 1993a,b; Janin et al, 1993).

- 4. Same as 3, but using video, rather than optical, viewing of the "outside" world. The difference between Classes 2 and 4 is that with 4 the displayed world should correspond orthoscopically with the immediate outside real world, thereby creating a "video see-through" system (e.g. Edwards et al, 1993; Fuchs et al, 1993), analogous with the optical see-through of option 3.

- 5. Completely graphic display environments, either completely immersive, partially immersive or otherwise, to which video "reality" is added (e.g. Metzger, 1993).

- 6. Completely graphic but partially immersive environments (e.g. large screen displays) in which real physical objects in the user's environment play a role in (or interfere with) the computer generated scene, such as in reaching in and "grabbing" something with one's own hand (e.g. Kaneko et al, 1993; Takemura & Kishino, 1992).

In addition, other more inclusive computer augmented environments have been developed in which real data are sensed and used to modify users' interactions with computer mediated worlds beyond conventional dedicated visual displays (e.g. Ishii et al, 1993; Krüger, 1993; Wellner, 1993; Mackay et al, 1993).

As far as terminology goes, even though the term "Mixed Reality" is not in common use, the related term "Augmented Reality" (AR) has in fact started to appear in the literature with increasing regularity. As an operational definition of Augmented Reality, we take the term to refer to any case in which an otherwise real environment is "augmented" by means of virtual (computer graphic) objects, as illustrated in Figure 1. The most prominent use of the term AR in the literature appears to be limited, however, to the Class 3 types of displays outlined above (e.g. Feiner et al, 1993a,b; Caudell & Mizell, 1992; Janin et al, 1993). In the authors' own laboratories, on the other hand, we have adopted this same term in reference to Class 1 displays as well (Drascic et al, 1993; Milgram et al, 1993), not for lack of a better name, but simply out of conviction that the term Augmented Reality is quite appropriate for describing the essence of computer graphic enhancement of video images of real scenes. This same logic extends to Class 2 and 4 displays also, of course.

Class 5 displays pose a small terminology problem, since that which is being augmented is not some direct representation of a real scene, but rather a *virtual* world, one that is generated primarily by computer. In keeping with the logic used above in support of the term Augmented Reality, we therefore proffer the straightforward suggestion that such displays be termed *"Augmented Virtuality"* (AV), as depicted in Figure 1º. Of course, as technology progresses, it may eventually become less straightforward to perceive whether the primary world being experienced is in fact predominantly "real" or predominantly "virtual", which may ultimately weaken the case for use of both *AR* and *AV* terms, but should not affect the validity of the more general *MR* term to cover the "grey area" in the centre of the virtuality continuum.

We note in addition that Class 6 displays go beyond Classes 1, 2, 4 and 5, in including directly viewed real-world objects also. As discussed below, the experience of viewing one's own *real* hand directly in front of one's self, for example, is quite distinct from viewing an image of the same real hand on a monitor, and the associated perceptual issues (not discussed in this paper) are also rather different. Finally, an interesting alternative solution to the terminology problem posed by Class 6 as well as composite Class 5 AR/AV displays might be the term*"Hybrid Reality"* (HR)ººº, as a way of encompassing the concept of blending many types of distinct display media.

º *Quoted from Call for Papers for this IEICE Transactions on Information Systems special issue on Networked Reality.*
ºº *Cohen (1993) has considered the same issue and proposed the term "Augmented Virtual Reality." As a means of maintaining a distinction betweem this class of displays and Augmented Reality, however, we find Cohen's terminology inadequate.*
ººº *One potential piece of derivative jargon which immediately springs to mind as on extension of the proposed term "Hybrid Reality" is the possibility that (using a liberal dose of poetic licence) we might refer to such displays as "Hyberspace"!*

---

2. The Need for a Taxonomy

The preceding discussion was intended to introduce the concept of Mixed Reality and some of its various manifestations. All of the classes of displays listed above clearly share the common feature of juxtaposing "real" entities together with "virtual" ones; however, a quick review of the sample classes cited above reveals, among other things, the following important distinctions:

- Some systems {1,2,4} are primarily video based and enhanced by computer graphics whereas

others {5,6} are primarily computer graphic based and enhanced by video.

- In some systems {3,6} the real world is viewed directly (through air or glass), whereas in others {1,2,4,5} real-world objects are scanned and then resynthesised on a display device (e.g. analogue or digital video).

- From the standpoint of the viewer relative to the world being viewed, some of the displays {1} are exocentric (WoW monitor based), whereas others {2,3,4,6} are egocentric (immersive).

- In some systems {3,4,6} it is imperative to maintain an accurate 1:1 orthoscopic mapping between the size and proportions of displayed images and the surrounding real-world environment, whereas for others {1,2} scaling is less critical, or not important at all.

Our point therefore is that, although the six classes of MR displays listed appear at first glance to be reasonably mutually delineated, the distinctions quickly become clouded when concepts such as real, virtual, direct view, egocentric, exocentric, orthoscopic, etc. are considered, especially in relation to implementation and perceptual issues. The result is that the different classes of displays can be grouped differently depending on the particular issue of interest. Our purpose in this paper is to present a taxonomy of those principal aspects of MR displays which subtend these practical issues.

The purpose of a taxonomy is to present an ordered classification, according to which theoretical discussions can be focused, developments evaluated, research conducted, and data meaningfully compared. Four noteworthy taxonomies in the literature which are relevant to the one presented here are summarised in the following.

- Sheridan (1992) proposed an operational measure of *presence* for remotely performed tasks, based on three determinants: extent of sensory information, control of relation of sensors to the environment, and ability to modify the physical environment. He further proposed that such tasks be assessed according to task difficulty and degree of automation.

- Zeltzer (1992) proposed a three dimensional taxonomy of *graphic simulation systems*, based on the components autonomy, interaction and presence. His "AIP cube" is frequently cited as a framework for categorising virtual environments.

- Naimark (1991a,b) proposed a taxonomy for categorising different approaches to recording and reproducing visual experience, leading to *realspace imaging*. These include: monoscopic imaging, stereoscopic imaging, multiscopic imaging, panoramics, surrogate travel and real-time imaging.

- Robinett (1992) proposed an extensive taxonomy for classifying different types of technologically mediated interactions, or *synthetic experience*, associated exclusively with HMD based systems. His taxonomy is essentially nine dimensional, encompassing causality, model source, time, space, superposition, display type, sensor type, action measurement type and actuator type. In his paper a variety of well known VR-related systems are classified relative to the proposed taxonomy.

Although the present paper makes extensive use of ideas from Naimark and the others cited, it is in many ways a response to Robinett's suggestion (Robinett, 1992, p. 230) that his taxonomy serve as "a starting point for discussion". It is important to point out the differences, however. Whereas technologically mediated experience is indeed an important component of our taxonomy, we are not focussing on the same question of how to classify different varieties of such interactions, as does Robinett's classification scheme. Our taxonomy is motivated instead, perhaps more narrowly, by the need to distinguish among the various technological requirements necessary for realising, and researching, mixed reality displays, with no restrictions on whether the environment is supposedly

immersive (HMD based) or not.

It is important to point out that, although we focus in this paper exclusively on mixed reality *visual* displays, many of the concepts proposed here pertain as well to analogous issues associated with other display modalities. For example, for auditory displays, rather than isolating the participant from all sounds in the immediate environment, by means of a helmet and/or headset, computer generated signals can instead be mixed with natural sounds from the immediate real environment. However, in order to "calibrate" an *auditory augmented reality* display accurately, it is necessary carefully to align binaural auditory signals with synthetically spatialised sound sources. Such a capability is being developed by Cohen and his colleagues, for example (Cohen et al, 1993), by convolving monaural signals with left/right pairs of directional transfer functions. Haptic displays (that is, information pertaining to sensations such as touch, pressure, etc.) are typically presented by means of some type of hand held master manipulator (e.g. Brooks, et al, 1990) or more distributed glove type devices (Shimoga, 1992). Since synthetically produced haptic information must in any case necessarily be superimposed on any existing haptic sensations otherwise produced by an actual physical manipulator or glove, *haptic AR* can almost be considered the natural mode of operation in this sense. *Vestibular AR* can similarly be considered a natural mode of operation, since any attempt to synthesise information about acceleration of the participant's body in an otherwise virtual environment, as is commonly performed in commercial and military flight simulators for example, must necessarily have to contend with existing ambient gravitational forces.

---

3. Distinguishing Virtual from Real: Definitions

Based on the examples cited above, it is obvious that as a first step in our taxonomy it is necessary to make a useful distinction between the concept of *real* and the concept of *virtual*. Our need to take this as a starting point derives from the simple fact that these two terms comprise the foundation of the now ubiquitous term "Virtual Reality". Intuitively this might lead us simply to define the two concepts as being orthogonal, since at first glance, as implied by Figure 1, the question of whether an object or a scene is real or virtual would not seem to be difficult to answer. Indeed, according to the conventional sense of VR (i.e. for completely virtual immersive environments), subtle differences in interpreting the two terms is not as critical, since the basic intention there is that a "virtual" world be synthesised, by computer, to give the participant the impression that that world is not actually artificial but is "real", and that the participant is "really" present within that world.

In many MR environments, on the other hand, such simple clarifications are not always sufficient. It has been our experience that discussions of Mixed Reality among researchers working on different classes of problems very often require dealing with questions such as whether particular objects or scenes being displayed are real or virtual, whether images of scanned data should be considered real or virtual, whether a real object must look 'realistic' whereas a virtual one need not, etc. For example, with Class 1 AR systems there is little difficulty in labelling the remotely viewed video scene as "real" and the computer generated images as "virtual". If we compare this instance, furthermore, to a Class 6 MR system in which one must reach into a computer generated scene with one's own hand and "grab" an object, there is also no doubt, in this case, that the object being grabbed is "virtual" and the hand is "real". Nevertheless, in comparing these two examples, it is clear that the reality of one's own hand and the reality of a video image are quite different, suggesting that a decision must be made about whether using the identical term "real" for both cases is indeed appropriate.

Our distinction between real and virtual is in fact treated here according to *three* different aspects, all illustrated in Figure 2. The first distinction is between *real objects* and *virtual objects*, both shown at the left of the figure. The operational definitions[o] that we adopt here are:

- Real objects are any objects that have an actual objective existence.

- Virtual objects are objects that exist in essence or effect, but not formally or actually.

In order for a real object to be viewed, it can either be observed directly or it can be sampled and then resynthesised via some display device. In order for a virtual object to be viewed, it must be *simulated*, since in essence it does not exist. This entails use of some sort of a description, or *model*[oo], of the object, as shown in Figure 2.

The second distinction concerns the issue of *image quality* as an aspect of reflecting reality. Large amounts of money and effort are being invested in developing technologies which will enable the production of images which look "real", where the standard of comparison for realism is taken as *direct viewing* (through air or glass) of a real object, or *"unmediated reality"* (Naimark, 1991a). *Non-direct viewing* of a real object relies on the use of some imaging system first to sample data about the object, for example using a video camera, laser or ultrasound scanner, etc., and then to resynthesise or reconstruct these data via some display medium, such as a (analogue) video or (digital) computer monitor. Virtual objects, on the other hand, by definition can not be sampled directly and thus can only be synthesised. Non-direct viewing of either real or virtual objects is depicted in Figure 2 as presentation via a Synthesising Display. (Examples of non-synthesising displays would be include binoculars, optical telescopes, etc., as well as ordinary glass windows.) In distinguishing here between direct and non-direct viewing, therefore, we are not in fact distinguishing real objects from virtual ones at all, since even synthesised images of formally non-existent virtual (i.e. non-real) objects can now be made to look extremely realistic. Our point is that just because an image "looks real" does not mean that the object being represented *is* real, and therefore the terminology we employ must be able carefully to reflect this difference.

Finally, in order to clarify our terms further, the third distinction we make is between real and virtual images. For this purpose we turn to the field of optics, and operationally define a *real image* as any image which has some luminosity at the location at which it appears to be located. This definition therefore includes direct viewing of a real object, as well as the image on the display screen of a non-directly viewed object. A *virtual image* can therefore be defined conversely as an image which has no luminosity at the location at which it appears, and includes such examples as holograms and mirror images. It also includes the interesting case of a stereoscopic display, as illustrated in Figure 2, for which each of the left and right eye images on the display screen is a real image, but the consequent fused percept in 3D space is virtual. With respect to MR environments, therefore, we consider any virtual image of an object as one which appears *transparent*, that is, which does not occlude other objects located behind it.

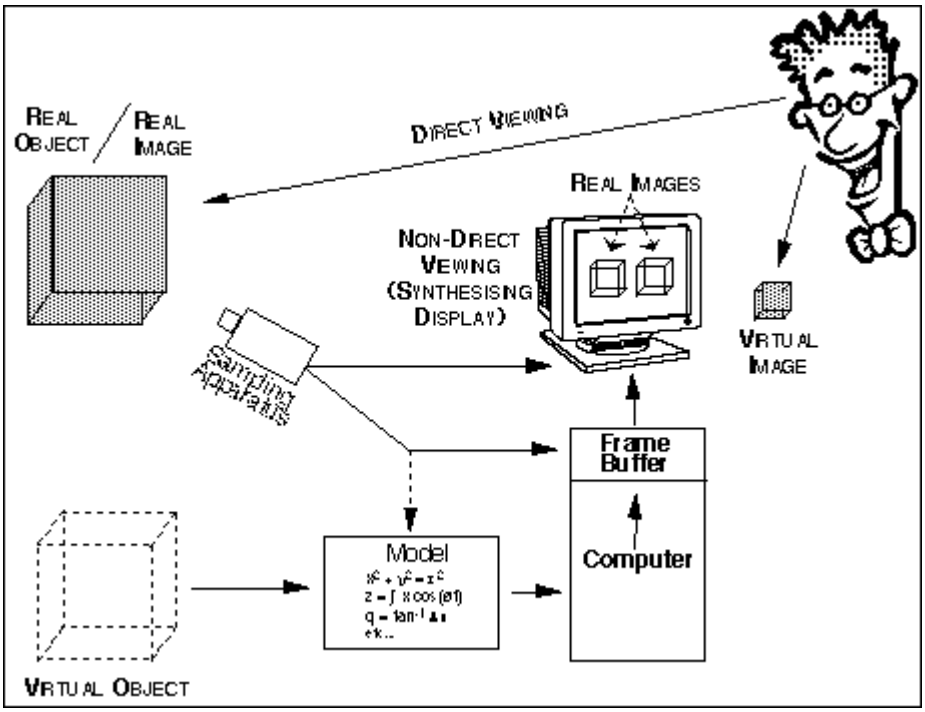

Figure 2: Different aspects of distinguishing *reality* from *virtuality*: i) Real vs Virtual Object; ii) Direct vs Non-direct viewing; iii) Real vs Virtual Image.

º *All definitions are consistent with the Oxford English Dictionary [30].*
ºº *Note that virtual objects can be designed around models of either non-exixtent objects or existing real objects, as indicated by the dashed arrow to the model in Fig. 2. A model of a virtual object can also be a real object itself of course, which is the case for sculptures, paintings, mockups, etc., however, we limit ourselves here to computer generated syntheses only.*

---

4. A Taxonomy for Merging Real and Virtual Worlds

In section 2 we presented a set of distinctions which were evident from the different Classes of MR displays listed earlier. The distinctions made there were based on whether the primary world comprises real or virtual objects, whether real objects are viewed directly or non-directly, whether the viewing is exocentric or egocentric, and whether or not there is an orthoscopic mapping between the real and virtual worlds. In the present section we extend those ideas further by transforming them into a more formalised taxonomy, which attempts to address the following questions:

- How much do we know about the world being displayed?
- How realistically are we able to display it?
- What is the extent of the illusion that the observer is present within that world?

As discussed in the following, the dimensions proposed for addressing these questions include respectively *Extent of World Knowledge, Reproduction Fidelity,* and *Extent of Presence Metaphor.*

4.1 Extent of World Knowledge

To understand the importance of the Extent of World Knowledge (EWK) dimension, we contrast this to the discussion of the Virtuality Continuum presented in Section 1, where various *implementations* of Mixed Reality were described, each one comprising a different proportion of real objects and

virtual objects within the composite picture. The point that we wish to make in the present section is that simply counting the relative number of objects, or proportion of pixels in a display image, is not a sufficiently insightful means for making design decisions about different MR display technologies. In other words, it is important to be able to distinguish between design options by highlighting the differences between underlying basic prerequisites, one of which relates to how much we know about the world being displayed.

To illustrate this point, in a paper by Milgram et al (1991) a variety of capabilities are described about the authors' display system for superimposing computer generated stereographic images onto stereovideo images (subsequently dubbed ARGOS™, for Augmented Reality through Graphic Overlays on Stereovideo (Drascic et al, 1993; Milgram et al, 1993)). Two of the capabilities described there are:

- a virtual stereographic pointer, plus tape measure, for interactively indicating the locations of real objects and making quantitative measurements of distances between points within a remotely viewed stereovideo scene;

- a means of superimposing a wireframe outline onto a remotely viewed real object, for enhancing the edges of that object, encoding task information onto the object, and so forth.

Superficially, in terms of simple classification along a Virtuality Continuum, there is no difference between these two cases; both comprise virtual graphic objects superimposed onto an otherwise completely video (real) background. Further reflection reveals an important fundamental difference, however. In that particular implementation of the virtual pointer / tape measure, the "loop" is closed by the human operator, whose job is to determine where the virtual object (the pointer) must be placed in the image, while the computer which draws the pointer has *no knowledge* at all about what is being pointed at. In the case of the wireframe object outline, on the other hand, two possible approaches to achieving this can be contemplated. By one method, the operator would interactively manipulate the wireframe (with 6 degrees of freedom) until it coincides with the location and attitude of the object, as she perceives it – which is fundamentally no different from the pointer example. By the other method, however, the computer would already *know* the geometry, location and attitude of the object relative to the remote cameras, and would place the wireframe onto the object.

The important fundamental difference between these sample cases, therefore, is the amount of knowledge held by the display computer about object shapes and locations within the two global worlds being presented. It is this factor, Extent of World Knowledge (EWK), rather than just accounting of the classes of objects in the MR mixture, that determines many of the operational capabilities of the display system. This dimension is illustrated in Figure 3, where it has been broken down into three main divisions.

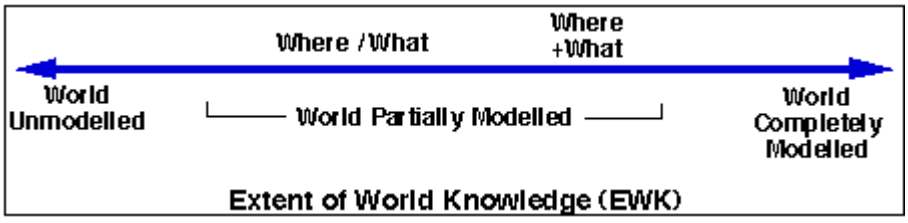

Figure 3: Extent of World Knowledge (EWK) dimension

At one extreme, on the left, is the case in which nothing is known about the (remote) world being displayed. This end of the continuum is reserved for images of objects that have been 'blindly' scanned and synthesised for non-direct viewing, as well as for directly viewed real objects. In the former instance, even though such an image might be displayed by means of a computer, no information is present within the knowledge base about the *contents* of that image. The other end of the EWK dimension defines the conditions necessary for displaying a completely virtual world, in

the 'conventional' sense of VR, which can be created only when the computer has complete knowledge about each object in that world, its location within that world, the location and viewpoint of the observer within that world and, when relevant, the viewer's attempts to change that world by manipulating objects within it.

The most interesting section of the EWK continuum of course is the portion which covers all cases between the two extrema, and the extent to which real and virtual objects can be merged into the same display will be shown to depend highly on the EWK dimension. In Figure 3, three types of subcases have been shown. The first, *"Where"*, refers to cases in which some quantitative data about locations in the remote world are available. For example, suppose the operator of a telerobot views a closed circuit video monitor and moves a simple cross-hair (a form of augmented reality) to a particular location on the screen. This action explicitly communicates to the computer that there is 'something of interest' at point {x,y} on that video image (or at point {x,y,z} if the cursor can be calibrated and manipulated in three dimensions), but it does not provide any enlightenment at all about *what* is at that location. Another illustration involves the processing of raw scanned data, obtained by means of video, laser, sonar, ultrasound scanners, etc., which on their own do not add any information at all about what or where objects in the scanned world are located. If, however, such data were to be passed through some kind of digital edge detection filters, for example, then the system could now be considered to have been taught some quantitative "where" type information, but nothing which would allow identification of what objects the various edges belong to.

The *"What"* label in Figure 3 refers to cases in which the control software does have some knowledge about objects in the image, but has no idea where they are. Reflecting a common case for many AR systems, suppose for example that, as part of a registration procedure, an accurate geometrical model of a calibration object is available. An image of that object can then be drawn graphically and superimposed upon an associated video image; however, unless the computer knows exactly where the real object is located and what its orientation is, in order to determine the correct scale factor and viewpoint, the two will not coincide, and the graphic object will appear simply to be floating arbitrarily within the rest of the remote scene.

Medical imaging is an important instance of an environment in which many of these factors are relevant. Many medical imaging systems are highly specialised and have as their objective the creation of a completely modelled world. For example, a system developed especially for cardiac imaging might perform model based fitting of raw scanned data, to generate a properly scaled image of the patient's cardiac system . If the scanned and reconstructed medical data were then to be superimposed upon a (video) image of the patient whence the data were taken, as in Fuchs et al (1993), the computer would have to have a model of not only how the reconstructed sampled data relate to each other, but also where corresponding points are located with respect to the real world, if accurate unmediated superimposition is to be possible.

4.2 Reproduction Fidelity

The remaining two dimensions both attempt to deal with the issue of *realism* in MR displays, but in diferent ways: in terms of *image quality* and in terms of immersion, or presence, within the display. It is interesting to note that this approach is somewhat different from those taken by others. Both Sheridan's (1992) and Robinett's (1992) taxonomies, for example, focus on the feeling of presence as the ultimate goal. This is consistent as well with the progression in "realspace imaging" technology outlined in Naimark's (1991a,b) taxonomy, towards more and more realistic displays which eventually make one feel that one is participating in "unmediated reality". In our taxonomy we purposely separate these two dimensions, however, in recognition of the practical usefulness of some high quality visual displays which nevertheless do not attempt to make one feel *within* the remote environment (e.g. Class 1), as well as some display situations in which the viewer in fact is already physically immersed within the displayed environment but may be provided with only relatively low quality graphical aids (e.g. Class 3 and 4).

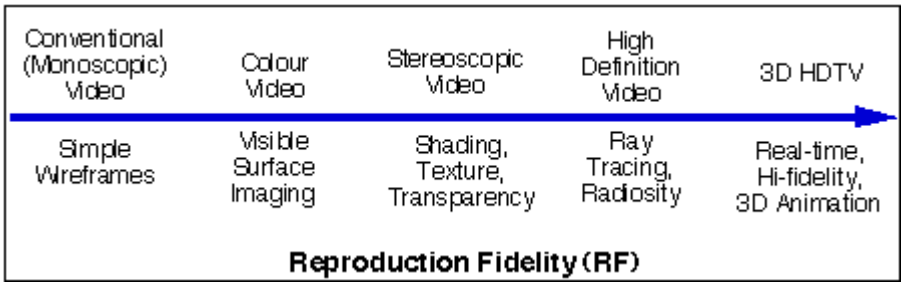

Figure 4: Reproduction Fidelity (RF) dimension.

The elements of the Reproduction Fidelity (RF) dimension are illustrated in Figure 4, where we follow the approach introduced in Figure 2 for classifying non-direct viewing, of either real objects or virtual objects. The term "Reproduction Fidelity" therefore refers to the quality with which the synthesising display is able to reproduce the actual or intended images of the objects being displayed. It is important to point out that this figure is actually a gross simplification of a complex topic, and in fact lumps together several different factors, such as display hardware, signal processing, graphic rendering techniques, etc., each of which could in turn be broken down into its own taxonomic elements.

In terms of our earlier discussion, it is important to realise that this dimension pertains to reproduction fidelity of *both* real and virtual objects. The reason for this is not only because many of the hardware issues are related. Even though the simplest wireframe display of a virtual object and the lowest quality video image of a real object are quite distinct, the converse is not true for the upper extrema. In Figure 4 the progression above the axis is meant to show a rough progression, mainly in hardware, of video reproduction technology. Below the axis the progression is towards more and more sophisticated computer graphic modelling and rendering techniques. At the right hand side of the figure, the "ultimate" video display, denoted here as "3D HDTV" might be just as close in quality to photorealism, or even direct viewing, as the 'ultimate' graphic rendering, denoted here as "real-time, hi-fidelity 3D animation". If this claim is accepted, one can then easily appreciate the problem if real and virtual display quality were to be treated as separate orthogonal dimensions, since if the maxima of each were ever reached, there would be no qualitative way for a human observer to distinguish between whether the image of the object or scene being displayed has been generated by means of data sampling or whether it arises synthetically from a model.

4.3 Extent of Presence Metaphor

The third dimension in our taxonomy, outlined in Figure 5, is the Extent of Presence Metaphor (EPM) axis, that is, the extent to which the observer is intended to feel "present" within the displayed scene. In including this dimension we recognise the fact that Mixed Reality displays include not only highly immersive environments, with a strong presence metaphor, such as Class 2, 3, 4 and 6 displays, but also important exocentric Class 1 type AR displays.

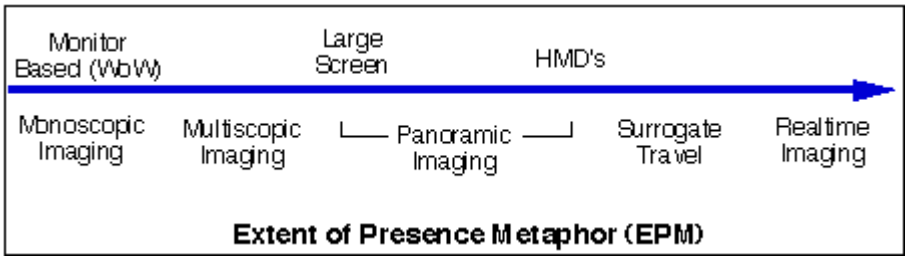

Figure 5: Extent of Presence Metaphor (EPM) dimension

Just as the subdimensions of RF for virtual and real objects in Section 4.2 were shown to be not quite orthogonal, so too is the EPM axis in some sense not entirely orthogonal to the RF axis, since each dimension independently tends towards an extremum which ideally is indistinguishable from

viewing reality directly. In the case of EPM the axis spans a range of cases extending from the metaphor by which the observer peers from outside into the world from a single fixed monoscopic viewpoint, up to the metaphor of "realtime imaging", by which the observer's sensations are ideally no different from those of unmediated reality. (Much of the terminology used in this section coincides with that used by Naimark in his proposed taxonomy of realspace imaging (Naimark, 1991a,b).) Along the top of the axis in Figure 5 is shown the progression of display media corresponding to the EPM cases below.

Adjacent to the monitor based class of WoW displays at the left hand side of the EPM axis are "Multiscopic Imaging" displays. These go beyond the class of stereoscopic displays indicated on the RF axis of Figure 4, to include displays which allow multiple points of view, that is, lateral movements of the observer's head while the body remains more or less still (Naimark, 1991a,b). The resulting sensation of local motion parallax should result in a much more convincing metaphor of presence than a simple static stereoscopic display. In order to accomplish multiscopic viewpoint dependent imaging, the observer's head position must normally be tracked. For simple scanned images (left hand side of the EWK axis in Figure 3) it is necessary to have either a sufficiently rapid and accurate remote head-slaved camera system or to be able to access or interpolate images within a rapid access video storage medium (e.g. Hirose, 1994; Liu & Skerjanc, 1992). Since the viewing of virtual world images (right hand side of the EWK axis in Figure 3), on the other hand, is less dependent on critical hardware components beyond reliable head-tracking, realisation of multiscopic imaging is somewhat more straightforward. Ware et al (1993) refer to such a display capability as "fish tank virtual reality".

Panoramic imaging is an extension of multiscopic imaging which allows the observer also to look around a scene, but based on the progressively more immersive metaphor of being on the inside, rather than on the outside looking in (Naimark, 1991a,b). Panoramic imaging can be realised partially by means of large screen displays, but the resulting metaphor is valid only for execution of tasks which are constrained to a suitably restricted working volume. A more inclusive instantiation of this class of displays can be realised by means of head-mounted displays (HMD's), which are inherently compatible with the metaphor of being on the inside of the world being viewed. Some of the technical issues associated with realising such displays are similar to those outlined above for multiscopic imaging.

Surrogate travel refers to the ability to move about within the world being viewed, while realtime imaging refers to the solution of temporally related issues, such as sufficiently rapid update rates, simulation of dynamics, etc. (Naimark, 1991a,b) The ultimate goal of "unmediated reality", not shown in Figure 5, should be indistinguishable from direct-viewing conditions at the actual site, either real or virtual.

---

5. Conclusion

In this paper we have defined the term "Mixed Reality", primarily through non-exhaustive examples of existing display systems in which real objects and virtual objects are displayed together. Rather than relying on obvious distinctions between the terms "real" and "virtual", we have attempted to probe deeper and examine some of the essential factors which distinguish different Mixed Reality display systems from each other: Extent of World Knowledge (EWK), Reproduction Fidelity (RF) and Extent of Presence Metaphor (EPM). One of our main objectives in presenting our taxonomy has been to clarify a number of terminology issues, in order that apparently unrelated developments being carried out by, among others, VR developers, computer scientists and (tele)robotics engineers can now be placed within a single framework, which will allow comparison of the essential similarities and differences between various research endeavours.

---

6. Acknowledgements

The authors gratefully acknowledge the generous support and contributions of Dr. N. Terashima, President of ATR Communication Systems Research Laboratories and Dr. K. Habara, Executive Vice President of ATR International (Chairman of the Board of ATR Communication Systems Research Laboratories). We also thank Mr. Karan Singh for his many helpful suggestions.

---

7. References

Bajura, M., Fuchs, H., Ohbuchi, R. (1992). Merging virtual objects with the real world: Seeing ultrasound imagery within the patient. *Computer Graphics, 26(2)*.

Brooks, FP Jr, Ming, O-Y, Batter, JJ, & Kilpatrick, PJ (1990). Project GROPE – Haptic displays for scientific visualization. *Computer Graphics 24(4),* 177-185.

Caudell, TP & Mizell, DW (1992). Augmented reality: An application of heads-up display technology to manual manufacturing processes. *Proc. IEEE Hawaii International Conference on Systems Sciences.*

Cohen, M. (1993). Besides immersion: Overlaid points of view and frames of reference. In *French-Japanese Workshop on Synthetic Worlds,*. Aizu-Wakamatsu, Japan: Wiley & Sons.

Cohen, M, Aoki, S & Koizumi, N (1993). Augmented audio reality: Telepresence/VR hybrid acoustic environments. *Proc. IEEE International workshop on robot and human communication (RO-MAN'93),* Tokyo, 361-364.

Drascic D, Grodski J, Milgram P, Ruffo K, Wong P, Zhai S (1993). ARGOS: A Display System for Augmenting Reality, *ACM SIGGRAPH Tech Video Review, Vol 88: InterCHI '93 Conf on Human Factors in Computing Systems*, (Abstract in *Proceedings of InterCHI'93*, p 521), Amsterdam, April 1993.

Edwards, EK, Rolland, JP & Keller, KP (1993). Video see-through design for merging of real and virtual environments. *Proc. IEEE Virtual Reality International Symposium (VRAIS'93),* Seattle, WA, 223-233.

Ellis, SR & Bucher, UJ (1992). Depth perception of stereoscopically presented virtual objects interacting with real background patterns. *Psychonomic Society Conference*, St. Louis, 1992.

Feiner, S, MacIntyre, B & Seligmann, D (1993). Knowledge-based augmented reality. *Communications of the ACM, 36(7),* 52-62.

Feiner, S, MacIntyre, B, Haupt, M & Solomon, E (1993). Windows on the world: 2D windows for 3D augmented reality. *Proc. ACM Symposium on User Interface Software and Technology (UIST'93)*, Atlanta, GA.

Fuchs, H., Bajura, M., Ohbuchi, R. (1993). Merging virtual objects with the real world: Seeing ultrasound imagery within the patient. *Video Proceedings of IEEE Virtual Reality International Symposium (VRAIS'93),* Seattle, WA..

Hirose, M., Takahashi, K., Koshizuka, T., Watanabe, Y. (1994). A study on image editing technology for synthetic sensation. *Proceedings of ICAT '94*, Tokyo.

Ishii, H., Kobayashi, M., and Grudin, J. (1993). Integration of interpersonal space and shared workspace: Clearboard design and experiments. *ACM Transactions on Information Systems (TOIS) (Special issue on CSCW'92), 11(4)*.

Janin, AL, Mizell, DW & Caudell, TP (1993). Calibration of head-mounted displays for augmented reality. *Proc. IEEE Virtual Reality International Symposium (VRAIS'93),* Seattle, WA, 246-255.

Kaneko, M., Kishino, F., Shimamura, K., Harashima, H. (1993). Toward the new era of visual communication. *IEICE Transactions on Communications,* Vol. E76-B(6), 577-591, June 1993.

Krüger, M (1993). Environmental technology: Making the real world virtual. *Communications of the ACM, 36(7),* 36-51.

Lion, D, Rosenberg, C & Barfield, W (1993). Overlaying three-dimensional computer graphics with stereoscopic live motion video: Applications for virtual environments. *SID Conference Proceedings*, 1993.

Little, W, Fowler, HW & Coulson, J / with revision by CT Onions (1988). *Oxford English Dictionary, Third Edition*. Clarendon Press: Oxford.

Liu, J. & Skerjanc, R. (1992). Construction of intermediate pictures for a multiview 3D system. *SPIE Proc. 1669, Stereoscopic Displays and Applications III*.

Mackay, W, Velay, G, Carter, K, Ma, C & Pagani, D (1993). Augmenting reality: Adding computational dimensions to paper. *Communications of the ACM, 36(7),* 96-97.

Metzger, PJ (1993). Adding reality to the virtual. *Proc. IEEE Virtual Reality International Symposium (VRAIS'93),* Seattle, WA, 7-13.

Milgram, P., Zhai, S., Drascic, D., & Grodski, J.J. (1993). Applications of augmented reality for human-robot communication. In *Proceedings of IROS'93: International Conference on Intelligent Robots and Systems*. Yokohama, Japan, 1467-1472.

Milgram, P, Drascic, D, Grodski, JJ (1991). Enhancement of 3-D video displays by means of superimposed stereographics. *Proceedings of Human Factors Society 35th Annual Meeting*, San Francisco, 1457-1461, Sept. 1991.

Naimark, M (1991a). Elements of realspace imaging. Apple Multimedia Lab Technical Report.

Naimark, M (1991b). Elements of realspace imaging: A proposed taxonomy. Proc. SPIE Vol. 1457, *Stereoscopic Displays and Applications II*.

Robinett, W (1992). Synthetic experience: A proposed taxonomy. *Presence, 1(2)*, 229-247.

Rosenberg, LB (1993). Virtual fixtures: Perceptual tools for telerobotic manipulation. *Proc. IEEE Virtual Reality International Symposium (VRAIS'93),* Seattle, WA, 76-82.

Sheridan, TB (1992). Musings on telepresence and virtual reality. *Presence 1(1),* 120-126.

Takemura, H & Kishino, F (1992). Cooperative work environment using virtual workspace. *Proc. Computer Supported Cooperative Work (CSCW'92)*, 226-232.

Tani, M, Yamaashi, K, Tanikohsi, K, Futakawa, M, Tanifuji, S (1992). Object-oriented video:

Interaction with real-world objects through live video. *Proc. CHI '92 Conf on Human Factors in Computing Systems*, 593-598.

Ware, C, Arthur, K & Booth, KS (1993). Fish tank virtual reality. *InterCHI '93 Conference on Human Factors in Computing Systems*, Amsterdam, 37-42.

Wellner, P (1993). Interacting with paper on the digital desk. *Communications of the ACM, 36(7),* 86-96.

Zeltzer, D (1992). Autonomy, interaction and presence. *Presence 1(1),* 127-132.