Proc. SPIE Vol. 2653: Stereoscopic Displays and Virtual Reality Systems III, San Jose, California, Feb. 1996. 123-134.

Perceptual Issues in Augmented Reality

David Drascic and Paul Milgram

ETC-Lab: Ergonomics in Teleoperation and Control Laboratory
Department of Mechanical and Industrial Engineering,
University of Toronto, Toronto, Ontario, Canada M5S 3G9
drascic@ie.utoronto.ca, milgram@ie.utoronto.ca, http://vered.rose.utoronto.ca/

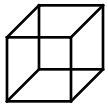
Abstract

Between the extremes of real life and Virtual Reality lies the spectrum of *Mixed Reality*, in which views of the real world are combined in some proportion with views of a virtual environment. Combining direct view, stereoscopic video, and stereoscopic graphics, Augmented Reality describes that class of displays that consists primarily of a real environment, with graphic enhancements or augmentations. Augmented Virtuality describes that class of displays that enhance the virtual experience by adding elements of the real environment. All Mixed Reality systems are limited in their capability of accurately displaying and controlled all relevant depth cues, and as a result, perceptual biases can interfere with task performance. In this paper we identify and discuss eighteen issues that pertain to Mixed Reality in general, and Augmented Reality in particular.

Keywords: stereoscopic displays, augmented reality, virtual reality, depth perception, human factors

Introduction

The principal objective of stereoscopic displays is to facilitate accurate perception of three dimensional scenes, by providing a potentially powerful binocular cue to supplement whatever monoscopic cues are already present in the scene. However, because the human visual system is very robust and is able to work with only partial data, it is simple to create some sensation of depth. The difficulty lies in the creation of an *accurate* sense of depth. For example, a 2D display of the Necker cube is typically perceived as a regular cube rather than as a flat collection of line segments, despite the absence of any depth cues. However, the lack of depth information in this case results in uncertainty about the orientation of the cube.



The Necker Cube

"Perception is not determined simply by the stimulus patterns; rather it is a dynamic searching for the best interpretation of the available data. ... The senses do not give us a picture of the world directly; rather they provide evidence for checking hypotheses about what lies before us. Indeed, we may say that a perceived object *is* a hypothesis, suggested and tested by sensory data. The Necker cube is a pattern which contains no clue as to which of two alternative hypotheses is correct. The perceptual system entertains first one then the other hypothesis, and never comes to a conclusion, for there is no best answer. Sometimes the eye and brain come to wrong conclusions, and then we suffer hallucinations or illusions." [1, p11-12]

Gregory stresses the important role of hypothesis testing in perception. A corollary to this idea, therefore, is the importance of *consistency* among the various cues which we use to infer information about our visual world.

In natural vision, the human visual system infers size and depth information using a variety of depth cues. In the real world, depth cues are almost always in agreement with each other, so an accurate perception of depth is possible. With stereoscopic displays, however, technological limitations are usually such that only a small subset of available depth cues can be fully implemented by any one system. All of the other potential depth cues are either missing, or not under the direct control of the system. Uncontrolled depth cues can end up providing *false* depth information, which can lead to distorted perceptions. Research suggests that perceptual uncertainty decreases, however, as the number of consistent depth cues increases, to the limits of the perceptual system. Beyond that, adding more depth cues may provide more local detail, but will probably not affect overall accuracy a great deal. [1, 2]

In situations where depth cues conflict, on the other hand, the outcome is more uncertain. Depending on the particular combination of cues involved, there are several potential results that we have observed at different times, which can affect performance in different ways. For example, if the task is one of aligning a real object with a virtual one, performance over time is a measure of two things: accuracy and consistency. If one examines the mean and the variance of the distribution of trials over time, one will find one of the following four effects:

1) One cue takes precedence over another, and the weaker cue is ignored. In this case the alignment of real and virtual objects will be accurate, and as consistent as the limits of the display system will allow.

2) The depth information provided by the two cues is combined, resulting in an intermediate percept. In this case there will be a regular error or bias in an alignment task, so that the real object will be consistently offset from the virtual one. However, repeated trials will find this bias to be consistent, and the variance of the distribution may not increase.

3) The cue conflict cannot be resolved, which leads to a rivalrous situation in which first one dominates, and then another. This causes either an increase in uncertainty about the spatial relationships, or an increase in inaccuracy. If the former, then subjects may report that when the real and virtual objects are in close proximity, it is impossible to determine exactly where they are aligned. They know when the real object is too close, and they know when it is too far, but between those extremes there is no clear alignment. Often they will simply position the real object in the middle of this range and hope for the best. If the second case, subjects may be satisfied that the objects are "close enough", but there may be a large variability in their responses. Naive subjects generally fall into the second category, while subjects with some experience in examining the perceptual issues of stereoscopic displays tend to fall into the first category.

4) The cue conflict can be resolved in different ways, depending on a variety of factors, including the subject's personal preferences, experience, the conscious attention of the subject to the various depth cues, the context of the images, and so on. In this case, performance will be unstable in the long term, even though consistent behaviour may be observed in the short term.

It is important to emphasise that, if the viewer is observing display images only passively, then these perceptual biases, uncertainties, and inaccuracies are often unimportant. Well-designed stereoscopic movies, for example, are rarely orthoscopic, and can actually exploit depth distortions as a dramatic tool. However, when the viewer must *interact* with the images, to perform some sort of precision manipulation task for example, the commission of errors assumes more practical importance.

Our objective in this paper is to report on work in progress aimed at identifying, classifying and characterising the class of visual perceptual issues that are associated with stereoscopic Augmented and Mixed Reality displays. In the next section we define these terms, and illustrate some of the consequences of the perceptual issues in question. We then present a brief review of cues that are known to play a role in depth perception, followed by a compendium of issues that, on the basis of perceptual theory, are expected to have some influence on the perception of depth in stereo MR displays. Note that it is not our aim in this paper to *quantify* the extent of those visual effects. Our ongoing research is devoted to that goal.

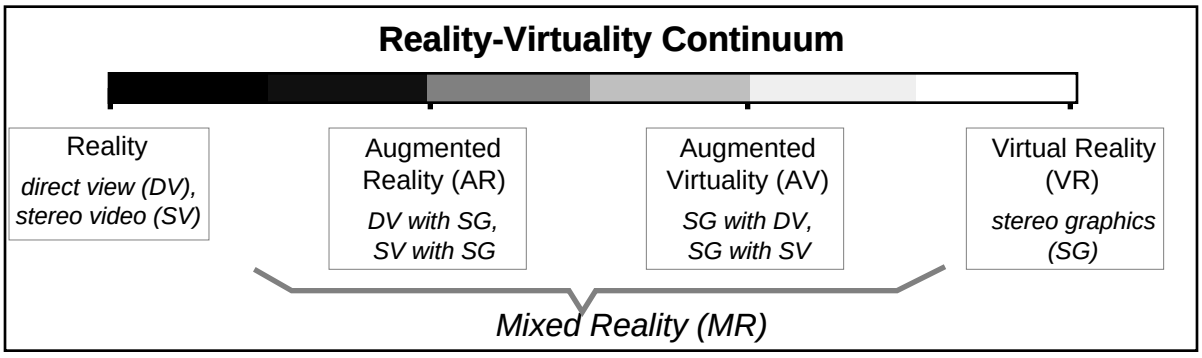


Figure 1: Simplified representation of the Reality-Virtuality Continuum, showing how real and virtual worlds can be combined in various proportions, according to the demands of different tasks.

Mixed and Augmented Reality

Definition: The definition and classification of "Augmented Reality" (AR), within the larger context of "Mixed Reality" (MR) displays, has been described in detail in [3]. In that paper, the concept of a Reality-Virtuality Continuum is introduced, to cover the spectrum of cases that define whether the world being viewed is primarily real or virtual. Real world objects can be displayed by simply scanning, transmitting and reproducing image data, as is the case with ordinary video displays, without the need for the display system to "know" anything about the objects. (This includes the viewing of real-world scenes either directly or via some optical medium.) On the other hand, *virtual* images can be produced only if the computer display system that generates the images has a model of the objects being drawn.

As illustrated in Figure 1, Mixed Reality refers to the class of all displays in which there is some combination a real environment and Virtual Reality. Within this context, the meaning of the term *Augmented Reality*, depicted on the left half of the continuum, becomes quite clear: AR displays are those in which the image is of a primarily real environment, which is enhanced, or augmented, with computer-generated imagery. Using a see-through head-mounted display, for example, it is

possible to make ghost-like images of anything we desire appear before the viewer in a fairly-well specified location in space. [4] These images can display information, or can serve as interactive tools for measuring or controlling the environment.

In contrast, *Augmented Virtuality* (AV) displays are those in which a virtual environment is enhanced, or augmented, through some addition of real world images or sensations. These additions could take the form of directly viewed (DV) objects, where the users might see their own bodies instead of computer-generated simulations, as is typical with VR. Augmented Virtuality could also combine VR with stereoscopic video (SV) images, where for example the view out of a virtual window might be of the real world at a distant location. *Augmented Virtual Tools* are an example of AV that were developed in order to solve a major limitation of VR, which is the absence of the sense of *touch.* In this system, real objects are fitted with special sensors and are used as the physical components of input devices for a VR system. The user "sees" virtual representations of these objects through the VR headset, which can have arbitrarily complex characteristics and functionality. [4] The illusion that the real object and the virtual one are the same thing is maintained as long as the displayed shape matches that of the real object, and its perceived size and location are approximately correct.

In this paper we assume that all display systems considered here are *stereoscopic.* Even though many of the display categories that we identify can be generalised beyond stereo displays, the perceptual issues we discuss are related primarily to stereoscopic displays.

Combining Multiple Display Modes: A survey of the literature reveals a wide range of display systems which have adopted the label of either Augmented Reality or Mixed Reality. Some systems, for example, combine monitor-based stereoscopic video images with superimposed stereoscopic graphics [5], while others use half-silvered mirrors and head-mounted displays to superimpose computer-generated images directly onto the real world [6, 7]. Still others use miniature cameras mounted on the viewer's head in order to combine SV and SG on a head-mounted display. [8, 9] Each of these technologies has its own particular applications, and each has its own benefits and problems.

In a general sense, there are three different primary display modes currently being used in Mixed Reality displays. The first is direct view (DV), where the user observes the real world directly, or by means of some optical device. The second encompasses stereoscopic video (SV) images of the real world. The third comprises stereoscopic graphic (SG) images of virtual objects and environments.

It is important to note that, as long as only one of these primary display modes is used on its own, any flaws in the presentation of depth information are usually manageable. When using SV for teleoperation, for example, only one mode is used to present the tools, the objects being manipulated, and the environment in which they are all situated. Any spatial distortions in the presentation of the remote scene will therefore affect all objects equally, and so in general it should be quite feasible to align a tool with an object accurately. [10] This kind of interaction is *indirect*, in that the user manipulates tools in the local environment, and the system interprets these actions and causes something to happen in the remote space. The user is unable to interact with the remote world directly.

However, if the user *is* able to interact *directly* with the images of the remote world, by reaching out and "touching" a displayed object, for example, then depth distortions in the display may cause the user to misperceive the relative locations of the real and displayed objects, and the user will have great difficulty manipulating those objects. This is an example of a mutli-modal MR system, that is, where different objects are presented using different kinds of display technologies. In such cases the depth distortions can be different for each mode, and objects that are designed to appear at the same depth may not be perceived as being at the same depth.

Research motivation: aligning real and virtual objects: The problem of misperceiving the location of objects at different depths is especially important if one of the principal functions of the viewer is to align two objects that are presented using different display modes. This problem can be readily observed with any large screen stereoscopic projection system, whether video or still photography. We have informally observed viewers using a variety of different stereoscopic projection systems. When the parallax between the left and right images is adjusted so that the image of an object appears within arm's reach of the subjects, and they are asked to quickly reach up and point to the bottom of that object and then hold their hand still in position, most people report that they have reached too far. When viewing the object only within the context of its particular display mode, they form one estimate of the position of that object. However, when they see the displayed object within the context of a different display mode, i.e. their own body, their understanding changes. Initially they perceive the object as being closer to the screen; however, when they can see both their hand and the object in close proximity, their disparity depth cue tells them that the object is in front of their hand. The farther the observer is away from the projection screen, the larger this error seems to be. (We are currently gathering empirical data to confirm this informal observation.)

When we ask our subjects to correct this alignment, most people bring their hand closer to themselves, and then spend a moment or two adjusting the depth of their hand until they reach a satisfactory alignment. Naive subjects generally feel satisfied that they have correctly judged the depth of the virtual object at this point, but observers trained in the perceptual issues of stereoscopic displays are often able to notice that some disparity remains, if their attention is drawn to this fact. If they are then asked to reduce the disparity to zero, they usually can do so by bringing their hand a little closer to their face, but invariably they complain that "it doesn't look right." This is probably due to accommodation differences between the two objects. In order to see their own hands clearly, subjects must accommodate, or focus, their eyes at the depth of their hands. In order to see the projected object clearly, however, subjects must accommodate at the depth of the display surface. Thus even when the retinal disparity between the real hand and virtual object is zero, it is impossible to see them both in focus at the same time. The accommodation depth cue in this case clearly tells the viewer that the two objects are at different depths.

Our observations of this phenomenon have led us to consider further the mechanism of this complex perceptual response. By identifying which depth cues are being used with each display mode, and by determining which depth cues are in conflict when more than one display mode is used, we have developed the framework presented below for discussing and analysing these issues. Our goal is to frame these problems in a useful ergonomic context, and thereby make researchers using multi-modal Mixed Reality displays aware of the perceptual issues at hand.

Overview of Depth Cues

As pointed out earlier in this paper, perception is an inferential process, and any given percept can be regarded as a hypothesis that is tested with evidence provided by the senses. [1] Almost all visual perception requires a sense of depth. Clearly our sense of size, orientation, and location requires a strong sense of depth, but even our sense of colour depends on our inference that the walls in the room we are looking at are angled in a particular way. As a result we perceive them to be painted in a single colour instead of the wide variety of shades of light that actually strike our eyes. [2]

Our understanding of the three dimensional nature of the environment around us is inferred from a variety of naturally occurring depth cues. In the real world (as opposed to the laboratory, or a virtual one), these cues are useful because they are consistently and reliably determined by the spatial arrangements of the objects that create them. These cues can be divided into four categories [2]: pictorial depth cues, kinetic depth cues, physiological depth cues, and binocular disparity cues.

Pictorial depth cues are those features in a flat picture that give the impression of objects being at various distances from the viewer. The strongest of these cues is *interposition*, where nearer objects hide more distant ones. Almost as important is *linear perspective*, where parallel lines appear to converge on the horizon. Linear perspective is based on perception of the contours of an object. A related depth cue is *texture (detail) perspective*, in which the density of a texture increases with distance. At greater distances, *aerial (atmospheric) perspective* indicates depth through the loss of detail, clarity, colouration, and a tendency towards a bluish-grey colour that can be observed with very distant objects.

Relative brightness is also a useful cue, in that objects farther away from a source of light appear darker than those closer to it, in the same way that a candle appears brighter when it is closer to us. Another good source of local detail about three dimensional relationships is *shadows*. Our brains automatically infer that a light source exists, generally overhead, and can extract a lot of spatial information from the way that shadows are cast from one object onto another.

Kinetic Depth Cues: While pictorial depth cues can be used to perceive depth in a flat picture, kinetic depth cues are those that provide depth information from changes in viewpoint, and movement in the objects being seen. *Relative motion parallax* refers to the way fixated distant objects seem to follow us as we move, while nearer ones appear to move in the opposite direction. The magnitude of these changes can provide a great deal of information about distances of those objects from each other, and from us. Similarly, *motion perspective*, such as when the raindrops near us appear to fall faster than those further away, is also a useful source of information. Whereas motion parallax is concerned with the relative movement of isolated objects, usually due to movement of the observer, motion perspective is concerned with whole gradients of motion that can occur whether the observer is moving or not.

The *kinetic depth effect* is another cue that that requires only the object to be moving, not the observer. One good practical use of this particular depth cue can be found in several data analysis programs for personal computers, which convey the structure of a three dimensional cloud of points by rotating that cloud about an axis perpendicular to the viewer's line of sight. Whenever the cloud is moving, its three-dimensional structure is quite apparent, but as soon as it stops, all depth is lost, and the cloud collapses back to a plane.

Physiological Depth Cues: In normally sighted humans, there are two related compensatory physiological systems that provide depth information. The first is that of *convergence*, where the two eyes rotate in opposite directions in order perceptually to fuse, or fixate on, an object. The relationship between convergence angle and distance is a learned response.

One stimulus for convergence is double vision, or *diplopia*. However, convergence responses can also be generated even when one eye is occluded, thanks to changes in the blurring of the retinal image of an object as it moves in depth.

Retinal blurring also drives *accommodation*, a change in the focal length of the eye. In order to perceive near objects clearly, the muscles attached to the lens of the eye must contract and pull it flatter, to reduce its optical power slightly. In order to see far objects clearly, the same muscles must relax to allow the lens to revert to its naturally more spherical state. The blurring of a retinal image is the primary drive for a change in accommodation, but is also driven to some extent by other depth cues, such as perspective, and by convergence. Accommodation serves as the second physiological depth cue.

It is important to note that the blur circles on the retina that stimulate both accommodation and convergence have also been shown to serve as a depth cue, whether convergence or accommodation change or not. For example, with very fast stimuli, when there is not enough time for any such responses, subjects are still able to make depth judgements in otherwise featureless situations. [2] However, for the sake of convenience, we shall lump together both the blur circles and the accommodative response in our definition of accommodation as a depth cue.

Binocular Disparity: In the same way that motion parallax reveals information about the arrangement of the environment through changing perspectives, having two eyes allows us to see the world from two different viewpoints without moving. By exploiting the differences, or *disparities*, between the left and right eye images, a strong sense of depth can be obtained. When our eyes fixate on a particular object, other objects that are closer to us are diplopic, with a retinal disparity in one direction, while objects that are further away are also diplopic, with a disparity in the other direction. The amount and direction of the disparity for each object serve collectively as a cue to how far in depth it is from the fixated object.

In order to use binocular retinal disparity as an absolute depth cue, however, it is necessary for the visual system to be able to scale the relationship between disparity and depth. When the fixation point is very near, the meaning of a certain angular disparity is very different from when the convergence point is very far. This scaling is done primarily using the convergence distance, but the accommodation distance is also relevant to some extent. [11]

The Perception of Size and Distance: In order to investigate the various perceptual biases associated with stereoscopic displays, it is important that researchers establish fair standards for comparison. The temptation might be to compare different performance and distance ratings obtained from these systems with the computed answers, based upon geometrical models. However, a more suitable comparison to make is perhaps not with some sort of "objective" correct answer, but instead with actual performance of related tasks by human operators under real world conditions. Decades of research into human perception of size and distance have indicated that both of these are very complicated mechanisms, which are affected by a range of issues much wider than just the depth cues presented above, and that even under the best of conditions, human perceptions can be quite different from "objective measures".

Perceptual Issues in Mixed and Augmented Reality

The following is a list of theoretical issues related to stereoscopic displays in general, with a particular interest in AR and MR displays. That is, some of the issues apply to any kind of stereoscopic display, while others are relevant only to mixed or virtual reality situations, or only to systems which incorporate viewpoint dependency through head tracking. No attempt has been made here to judge the various issues in terms of importance, severity, or priority. The issues discussed have been grouped into three categories: "implementation errors", which can be solved through careful application of currently available technology; "current technological limitations", which will presumably become less important as the state of the art improves; and "hard problems", that require new fundamental developments in technology to be solved.

In the discussion below, SV refers to stereoscopic video, SG refers to stereoscopic graphics, and DV refers to direct view. Display devices are categorised into one of three types: head-mounted displays (HMDs), desktop monitors, and large screen projection systems. Large screen projection systems in turn include both front and rear projection systems, which are designed to subtend a major portion of the user's visual field. Monitors on the other hand can come in a variety of sizes, but generally subtend only a small portion of the user's visual field. HMDs use small imaging devices and optics to enlarge the effective field of view and put the images on a particular accommodation plane. Devices such as the FakeSpace™ Boom are considered to be a HMD for the purposes of this discussion.

Implementation Errors

Calibration Errors: In order for a graphic or video image to be scaled properly in space, the calibration parameters which determine the visual angle, perspective, and binocular parallax of that image relative to the viewer must be accurately specified. While measuring the geometry of such systems is a solved problem [12], it is not a trivial one, especially for SV. It is extremely difficult, for example, to measure the location of one camera relative to the other in a purely physical way,

since the location of the nodal points of the camera lenses and of the imaging plane are not directly measurable. Such measurements must in general be done indirectly, by analysing the images of the two cameras, and using machine-vision techniques to determine their relative orientation. [5]

It is possible to design the camera and display system so that the stereoscopic image on a particular monitor, when viewed from a particular viewpoint, is *orthoscopic*, where all visual and disparity angles correspond with those in the real world. [13, 14] Only in this particular case does a stereoscopic image portray an accurately scaled rendition, where the scale is precisely 1:1 vertically, laterally, and in depth. However, once the viewer's head moves from the specified location, the visual space is again distorted. Truly orthoscopic systems are consequently hard to achieve and harder to maintain, and often the demands of the task make them undesirable, in that the field of view afforded by an orthoscopic system is often too small for useful work.

Depth distortions due to calibration errors can affect performance in many different ways. For example, hyper-stereoscopic systems, which create images that make it seem that the viewer's eyes are further apart than they really are, act to warp space by exaggerating depth near to the cameras while collapsing depth at a distance. This can be very useful for some kinds of tasks. One of the negative side effects, however, is that this creates the illusion that objects moving toward the camera at a constant speed appear to be accelerating. When remotely driving a vehicle to approach a target, for example, users find it difficult to estimate an accurate intersection time, and can easily overshoot the target, or collide with it. [15, 16] Similar depth distortions occur quite frequently with head mounted virtual reality systems. Unless both the display and the software are adjusted to exactly match the inter-pupillary separation of the user, the presented spatial relationships are always distorted. People *can* adapt rather quickly to mis-calibrated systems, by re-calibrating their own visual, vestibular, and proprioceptive systems. [17] This kind of adaptation can occur fairly rapidly. However, in situations where the viewer must switch back and forth between the real and virtual environments, these mismatches in co-ordinate systems can lead to disorientation and other performance deficits. [18]

Calibration Mismatches: In principle, augmented reality systems that combine graphics with video can be implemented in such a way that the graphics are drawn with exactly the same calibration parameters as the video. To accomplish this, the camera system parameters must be measured with sufficient accuracy for the graphics system to match them. When this is done, users are able to align graphic objects with video ones with a good degree of accuracy. [5, 19]

On the other hand, mixed reality systems that combine graphics with the directly viewed real world (e.g. see-through AR) are very sensitive to spatial distortions. If the graphics in this case are not absolutely orthoscopic, corresponding co-ordinates in the real and virtual worlds will not be aligned, and users will be unable to interact properly with their environment.

Interpupillary Distance Mismatches: The convergence point of an object in a stereoscopic viewing system depends on the interpupillary distance (IPD) of the observer, and unless the true IPD of the particular viewer is taken into account, the scale of the stereoscopic image in depth will be incorrect. Even small errors in IPD measurement can lead to large errors. [20] For example, if the system assumes the IPD of the subject is 65mm, whereas in fact the subject's true IPD is 64mm, and the subject is situated 80 cm away from the display screen, then an object intended to be displayed at a distance of 10.4 metres in front of the subject (i.e. behind the screen) will in fact be converged upon at 12.8 metres away. Furthermore, if the subject's true IPD is 61mm, the convergence point will be 48.8 metres away, a 369% error! If the subject's true IPD is 60mm, she will have to diverge her eyes in order to fuse the same image. We note here that human IPDs range from approximately 45mm to 75mm.

The convergence of the eyes also helps the visual system to scale both the disparity between the left and right eye images (i.e. the length of the object in depth), and the apparent height and width of the object being viewed. [1] Incorrect IPD adjustments can cause objects to be warped by different amounts in these three dimensions, so they might appear miniature but stretched in depth, for example.

If there is an error in IPD, but all objects are presented using a single display mode, then the magnitude of the perceived location error will be the same for both the graphic objects and the video objects (assuming correct calibration for other parameters), so there should be no relative mismatch errors. This means that slow alignment tasks need not be affected significantly. However, if the user is expected to move quickly and comfortably through such an environment, the differences in scaling between the visual world and the kinaesthetic / proprioceptive one will necessarily degrade performance. Even when we have very fast and responsive VR systems, for example, it will still be very difficult to perform complex psychomotor tasks such as playing virtual ping pong or racket ball until we can adjust the display to exactly match the IPD of the viewer. On the other hand, when a multi-modal interaction is required, such as between direct view and SG, then any errors in IPD estimation will cause immediate problems in grasping and alignment tasks.

Current Technological Limitations

Static and Dynamic Registration Mismatches: Assuming that a calibrated, orthoscopic graphic display can in fact be created for an augmented or mixed reality display system, the problem remains of how to *register*, or align, the co-ordinate system of the virtual world with that of the real world. Furthermore, even though the scales may match exactly, it is difficult to maintain a precise alignment between the graphics and the video, or between the graphics and the direct view, given the differential lags associated with the different display modes. For example, the visual changes associated with turning one's head are essentially instantaneous in direct view, while the lag in the visual response from a graphics or video system may be several tens or hundreds of milliseconds. Maintaining registration in a dynamic, changing environment is a very challenging technological issue. An AR system where the graphics lag the direct view can quickly lead to dizziness, nausea, and other symptoms similar to simulator sickness, which is caused by disparities between information received by the visual, vestibular, and proprioceptive senses. [21]

Restricted Field of View: A complete and accurate sense of space requires a very wide field of view. Our understanding of the world around us is generally built up, one piece at a time, using the relative arrangements of each item to help solidify our percepts. When, due to an unnaturally narrow field of view, we are unable to see important parts of the world, such as the floor and our body within the world, we lose a great deal of confidence in our understanding of the world. Simple actions, like walking around a corner, can be quite difficult when the user is wearing a head-mounted display with a very limited field of view. The larger the field of view, in general, the more complete and accurate depth perception will be.

Limitations and Mismatches of Resolution and Image Clarity: Objects that are displayed on a HMD, monitor, or projector necessarily have less resolution, or *spatial frequency*, than directly viewed objects. This means that correspondingly less texture is provided than a real object in the same position would possess. If this lack of detail is inappropriately interpreted as texture perspective, the imaged object may appear to be farther away than the real world object would appear.

In the ARTEMIS augmented reality system discussed elsewhere in this volume [5], for example, the SV camera images are captured at a resolution of 640 by 480 pixels, while the SG images are drawn at a resolution of 1280 by 480. Because the graphics have higher resolution, and since they are, in general, aliased, the edges of virtual objects appear much sharper than the edges of real objects, which are not aliased. [22] This difference in resolution and clarity could be interpreted by the brain as a difference in accommodation, which could in turn serve to make a video object appear further away than a graphic object, even when there is in fact no stereoscopic disparity between them.

Similarly, when viewing a virtual environment on a large screen projection display and attempting to interact with it using direct touch, the graphic images will look fuzzy and low resolution compared with the directly viewed objects, and viewers may thus perceive the graphic objects to be farther away than the real ones, even when the stereoscopic disparity between them is zero.

Luminance Limitations and Mismatches: Few displays are able to produce the range and intensity of luminance that is typically experienced in the real world. Acuity in general, and stereoacuity in particular, are known to be worse under low light conditions than under bright light conditions. [23] High luminance levels are very much desired in film projection, as being more pleasing and dramatic for the audience. IMAX has devoted a great deal of money in the development of the brightest possible motion picture projection system currently available. [24, 25] The effect of illumination on stereoscopic viewing can also be seen using a stereoscopic slide viewer, such as a Mattel ViewMaster. Using low level ambient indoor lighting as the illumination source of the stereo slides gives a sense of depth, and the image looks satisfactory. However, using a much more intense light source, such as direct sunlight or the intense light at the focal point of an overhead projector results in an image that is much more pleasing, with a much stronger sense of the viewed space, and a much greater experience of *presence*. [26] What effect this might have on performance is undoubtedly task dependent, but it is not unreasonable to expect some benefit from brighter displays.

In mixed reality situations involving direct view, hardware limitations can easily result in displayed images that are much less bright than directly viewed objects. Because brighter objects appear closer than less bright objects, the darker displayed images may once again appear to be farther away than the real objects, even when there is no stereoscopic disparity between the two objects. Such an effect could be expected to have significant consequences for tasks which require alignment of real and virtual objects, or spatial grasping of a virtual object by the viewer's hand.

Contrast Mismatches: Contrast ratio refers to the differential luminance between two adjacent points. The maximum contrast ratio of HMDs, monitors, and projection systems is far below the contrast ratios experienced with direct view on a bright, sunny day. Imaged objects, therefore, may have much less contrast than adjacent real object in a mixed reality system. Such low contrast may be inappropriately interpreted as aerial perspective, and the imaged objects may appear to be further away than they otherwise should.

Unless viewpoint adjustments are perfectly compensated for, the contrast ratio of imaged images may fluctuate as the observer moves around in space, depending on the particular technology involved. Such changes with position may also contribute towards differential depth errors with position.

Size & Distance Mismatches: Because differences in image resolution and clarity can be interpreted as accommodation and texture perspective depth cues, and because image brightness and contrast mismatches can be inappropriately interpreted as luminance and aerial perspective depth cues, there is the possibility that such false cues can cause otherwise identical video, graphical, and directly viewed objects to appear to have different sizes or to be at different distances, even when the geometry of the images suggests they this should not be so.

Fuzzy images cannot be properly accommodated, which can lead to uncertainty or misinterpretation of the size and depth of the images. If the object is familiar, or if its size is assumed by the observer to be known, then if its size appears to be smaller than it should, this will imply that it is farther than it should be, and vice versa. The effects of this could be quite significant, and can also vary depending on the display environment.

Limited Depth Resolution: The limitation of current display resolution affects not only the clarity and spatial frequency of displayed objects, but it also limits the precision with which binocular disparities can be perceived. Using typical, aliased graphics systems, for example, the minimum step for changes in disparity corresponds to a single pixel on the display. Using anti-aliased graphics can improve the disparity resolution to one fifth of a pixel [27], but even this can be very coarse depth resolution, especially in the space behind the display surface. This limited resolution can affect the performance of an alignment task, especially if the desired alignment distance is not exactly achievable.

Vertical Alignment Mismatches: In carefully designed video displays, it is important to exclude alignment errors, such as vertical disparities, between the left and right images. However, all converged SV camera systems will have some residual vertical disparities at the edges of the screen. Furthermore, the interlaced nature of an alternating-field SV image means that there is a one pixel vertical disparity between the left and right fields when viewed on an NTSC monitor. This vertical disparity is not very noticeable for non-aliased video images, but it shows up very clearly in the aliased graphics of an NTSC-based AR system, such as the ARGOS system developed in our lab. [15, 28] If the video cameras are carefully aligned to minimise vertical disparities, then there would still be a difference of approximately one pixel in the vertical registration of the graphics with respect to the video. Under most circumstances, this registration problem is not noticed, until the user must align a real and a virtual object. When the two objects are close together, the residual vertical disparity in the graphic object compared to the video one leads to a situation of uncertainty. However, as the two objects approach, there is a point at which one is clearly in front of the other, and there is another point where this relationship is clearly reversed. Between these two positions there is a region where the subject knows that the two objects are not aligned, but is powerless to improve the alignment. Most subjects attempt to place the pointer in the middle of that region and hope for the best, but this *zone of uncertainty* leads to greater variability in their alignment performance.

Similar effects can be expected in see-through HMDs used for AR. Should there be any misalignment at all between the orientation of the viewer's head and the orientation of the graphics, small vertical disparities will be apparent in the graphics that will not be present in the real world. If these vertical disparities are small enough, very few subjects will notice them; instead, most people will unconsciously compensate for them when looking at the graphics, and compensate again for their absence when looking at the real world. However, alignment tasks will be faced with a certain *zone of uncertainty*, which will degrade performance. This may have been a contributing factor to the larger-than-expected variability found in the AR alignment task described in [6].

Viewpoint Dependency Mismatches: In a well-designed mixed reality system, the graphic and video images should track the viewpoint of the observer accurately and without lag. From a mathematical, software, and graphics hardware point of view, this is a solved problem. From a *measurement* point of view, however, this is an unsolved problem, since there are few if any sufficiently accurate head trackers. Assuming that adequate head tracking is available, the hardware to slave the cameras to the head movement is likely to be very expensive, and there are also likely to be lags in the system Unless the lags of the graphics and the cameras are well matched, there is a likelihood for misalignments during any rapid head movement.

The Hard Problems

Interposition Failures: Interposition, or occlusion, is the strongest depth cue, in that it can override all others. One fact that dominates our perception of the world is that no matter how ambiguous the situation might otherwise be, near objects cannot be occluded by far objects. In rich, real world environments this is especially true, but even under reduced, controlled laboratory conditions, most subjects show a preference for interposition over the binocular cues. [29]

One major advantage of mixed reality systems is that only a partial computer model of the real world is required in order to impart useful functionality to the system. Acquiring and maintaining a model of the real environment can be a very difficult and expensive problem to solve. Without a complete model of that environment, however, it is impossible to fully implement the occlusion cue in mixed reality.

In AR systems using SV with SG, the graphics will always occlude the video, except when the world model indicates that it should not, should such a model be available. Using the *ARTEMIS* system developed in our lab, it is possible to interactively define a boundary or plane that corresponds with a real object. When virtual objects pass behind this plane, they appear to disappear smoothly into it, simulating interposition. Without this model, the virtual objects are drawn regardless of the layout of the real world. Often under these conditions, subjects report that the Virtual Pointer is a double image, and they cannot fuse the left and right halves. [22] For example, if the very simple image of the *Virtual Pointer* (a monochromatic line drawing of a downward pointing arrow) is moved slowly away from the viewer until it passes behind a textured wall, or into a piece of equipment, the viewer's knowledge of the properties of walls and equipment seems to take precedence, and the pointer becomes diplopic, or doubled, as is passes behind the surface, violating the interposition cue. It is possible that such diplopia can actually be used as a depth cue. [22]

The occurrence of this diplopia seems to be at least partly a function of the features of the objects being viewed. If the real object has a flat, featureless surface, for example, the doubled image of the virtual pointer occurs much less quickly, and seems to require a much larger violation of the interposition cue. Instead, the surface of the object is seen as being discontinuous, or sometimes as transparent. A similarly effect is found if the pointer is a more complex three dimensional pointer that is rotating around its vertical axis with a richly implemented lighting model. This pointer will be much more robust to moderate violations of interposition, even with richly featured real objects. Often, the real object becomes the diplopic one.

In AR systems using DV and SG, where the graphics are viewed through a half-silvered mirror, the graphics always appear transparent, and so never really occlude the real world, and they in turn are never occluded by the real world, except when a computer model indicates that they should be. In the absence of such a model, it is a easy to position a virtual object behind the surface of a real one. The results in this case can be quite complex. In a study by [6], they used an extremely simple virtual object (a monochromatic four-sided wire-frame pyramid that rotated very slowly about its vertical axis), a simply patterned real object (a checkerboard), and an LED pointer to measure the apparent depth of the virtual pyramid. In the absence of the real object, subjects were able to match the disparity of the virtual pyramid and the real LED with reasonable consistency, although with an unexplained bias and larger than expected variability. In the presence of the real object at the same distance as the indicated depth of the pyramid, the pyramid was judged by most subjects to have jumped forward in space, so as to be in front of the checkerboard. This is a logical assumption since the checkerboard did not occlude the pyramid. When the checkerboard was moved forward a large amount, so that convergence clearly indicated that it was in front of the pyramid, many subjects reported that the checkerboard appeared *transparent.* When the checkerboard rotated about the viewer's line of site, many subjects reported that the pyramid was unfusable.

The ARTEMIS system mentioned above presents a virtual simulated robot super-imposed on the SV image of the corresponding real world robot. The user controls the local simulation in the stereoscopic context of the remote (video) world. The virtual robot can be presented in a number of ways. Using a solid model of the robot can look very realistic, but it occludes the real world SV image too much. However, when a wire-frame outline of the robot is used instead, violations of occlusion are much less important, and users can successfully manipulate remote object using the virtual robot, even though no attempt to model interposition is made. [22]

Expanded Depth of Field: In direct view, diplopic images are typically out of focus, because of the restricted depth of field of the viewer. Stereoscopic displays, however, typically have an "infinite", or at least greatly expanded, depth of field. This can make the diplopic images much more distracting and interfering than they would be with direct view. [30] In the miniature stereoscopic cameras used for endoscopic surgery, depth of field is sometimes restricted, and at least one manufacturer claims that this reduces the distracting effect of large disparities in the image. [31] In general, however, SV systems tend to have very large depths of field, which suggests that there might be less tolerance for larger disparities.

Software techniques can be used to implement a simulated depth of field with graphical displays, which has proven quite successful in creating photo-realistic graphic images. In order for this to work with interactive SG displays, however, the computer would have to be able to adjust the image according the current convergence point of the user. This is theoretically possible, but challenging, and the benefits obtained through such an approach may not be worth the expense required.

Absence of Accommodation: Perception of depth and perception of apparent size in purely optical viewing systems show a consistent tendency to "minify" objects. This occurs with monocular and binocular systems (telescopes and binoculars), with see-through optics and with ground-glass imaging planes (photographic view-finders and projection

periscopes), and with unity magnification and higher power magnification. [11, 32, 33] While there are a number of contributing factors, and this effect is reduced in scenes that are rich in depth cues, a fixed, incorrect accommodation distance is a major part of the problem. Even when the objects are correctly perceived at the right depth, they are perceived as being a little bit smaller than they really are. [34] In all of these cases, the imaging plane is accommodated closer to the viewer than the objects being viewed would be.

Furthermore, the amounts that judgements of distance and judgements of size are affected are not necessarily the same, and are differentially affected by environmental factors, such as the availability of other depth cues. [34] These finding suggest that variations in viewing distance from the display surface may change a subject's perception of the size and location of virtual objects differently than the geometry may suggest, and the variability of the results presented in the aforementioned studies indicate that it will be hard to allow for, or control, these factors.

Accommodation - Vergence Conflict: All stereoscopic display systems require the viewer's eyes to accommodate (focus) at the depth of the display in order to see the objects clearly, regardless of the position of the objects. Accommodation and convergence are closely linked visual systems, and each drives the other under different conditions. Many writers assert that disturbing the accommodation-vergence system is the major cause of eye-strain in stereoscopic displays. None cite any research which actually demonstrates that this is the case, however. All optical systems, including binoculars, eye-glasses, microscopes, camera view-finders, and so on, disturb this system to some extent. It is known to be flexible and robust, and can adapt to new situations fairly quickly. When Cegalis measured accommodative responses to changes in convergence distance caused by ophthalmic prisms, he found that although subjects initially under-accommodated the target, they quickly adapted for near target distances, with no significant after-effects. [35]

Furthermore, the resting state of these two systems is vastly different. While accommodation and vergence are closely correlated for certain visual tasks, the systems that control them do not appear to strive to maintain equal distances. For example, if you cover one eye for a few seconds, the natural tendency is for that eye to quickly drift towards the nose. Brief accommodation variances are sometimes noticed, but they quickly pass without any conscious effort on the part of the viewer. [36] Furthermore, when Kersten and Legge presented subjects with speckle patterns, which cannot be accommodated by retinal blur, they found that convergence-driven accommodation and convergence are linearly related with a mean ratio of 0.9, not 1.0. This difference is consistent with other studies that used small pupils to prevent retinal blur from driving accommodation. [37]

Based on these reports, it is reasonable to be sceptical about unsupported claims regarding the painful and debilitating effects of accommodation-vergence conflict, especially in light of subjective reports indicating that stereoscopic displays with large parallax demands are at least as comfortable as, if not more so, comparable monoscopic displays, even for tasks lasting three hours or longer. [14] On the other hand, even if accommodation-vergence conflict is not a major contributor to eye strain, it may still be a major factor in perceptual errors of stereoscopic displays.

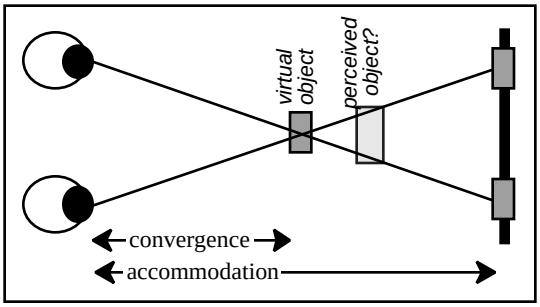
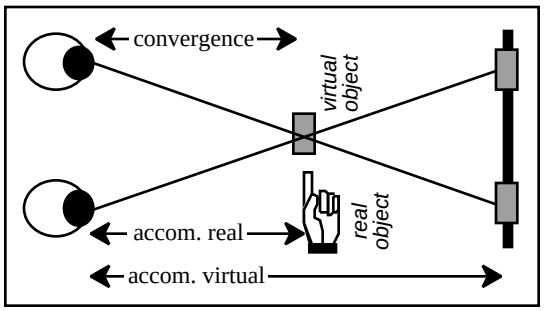


Figure 2: Effects of Accommodation - Vergence Conflict on the perceived size and position of virtual objects.

Most Virtual and Augmented Reality systems presume that perceived depth of an object is at the convergence point. However, as discussed above, this is not necessarily the case. While accommodation is a weak depth cue at best, it can certainly affect both the perception of depth and the perception of size.

As described anecdotally at the beginning of this paper, most people see virtual objects, like the one presented in Figure **2**, at a distance closer to the screen than convergence alone might suggest. This could be due in part to the accommodation-vergence conflict.

Accommodation Mismatch: In a mixed reality situation involving direct view, there will almost always be an accommodation mismatch between real and virtual objects. The accommodation distance for virtual objects is always the distance from the viewer to the display screen (or accommodation plane, for see-through AR), whereas the accommodation distance for the real object corresponds with the position of the object. When the viewer tries to align his finger with a virtual object, the differences in accommodation will provide an unmistakable depth cue, saying that the virtual object *must* be at a different depth than the real one. The greater the differences in accommodation distances, the stronger this cue is likely to be, and the more significant its impact on performance.



For see-through displays, it is possible to design dynamically compensating optics so that the accommodation depth always matches the convergence depth of the graphical object. However, if several different objects are presented, then the optics would have to track the convergence distance of the viewer's eyes in order to eliminate the effects of Accommodation - Vergence Conflict and Accommodation Mismatch.

Figure 3: Accommodation Mismatch between real and virtual objects tells the viewer that the objects must be at different distances, even when there is no disparity.

Absence of Shadow Cues: Regardless of how well-developed a model of the real world an AR system has, it is almost impossible for it to create realistic shadows that appear to fall on real objects, even for video-based AR systems, especially under complex, real world conditions. Shadows are critical for teleoperation using monoscopic video, and play a very important role even with stereoscopic video. Their absence from mixed reality systems can greatly impair performance in the field.

Conclusion

There are a wide variety of factors that conspire to impede the success of interactive systems that combine direct view, stereoscopic video, and virtual reality in some combination, by introducing misleading or conflicting depth cues. It is crucial that designers of such systems be aware of these problems, and design their systems around the weaknesses these perceptual issues can create.

Acknowledgements

Portions of this work were conducted while the authors were working in the ATR Communication Systems Research Laboratory, Kyoto, Japan, under the sponsorship of Dr. Fumio Kishino. The authors are indebted to ATR CSRL for its generous support.

References

1. Gregory, R.L., *Eye and Brain: The Psychology of Seeing*. 2nd ed. World University Library. 1973, Toronto: McGraw-Hill Book Company. 251.
2. Kaufman, L., *Perception: The World Transformed*. 1979, New York: Oxford University Press. 416.
3. Milgram, P., *et al. Augmented reality: a class of displays on the reality-virtuality continuum*. in *SPIE Volume 2351: Telemanipulator and Telepresence Technologies*. 1994.
4. Sowizral, H.A. *Interacting with virtual environments using augmented virtual tools*. in *Stereoscopic Displays and Virtual Reality Systems*. 1994. San Jose.
5. Rastogi, A., *et al. Telerobotic Control with Augmented Reality*. in *SPIE Volume 2653: Stereoscopic Displays and Applications VII*. 1996. San Jose.
6. Ellis, S.R. and U.J. Bucher. *Distance Perception of Stereoscopically Presented Virtual Objects Optically Superimposed on Physical Objects by a Head Mounted See-Through Display*. in *Proceedings of the 38th Annual Meeting of the Human Factors and Ergonomics Society*. 1994. Nashville, Tennesee.
7. Rolland, J.P., W. Gibson, and D. Ariely, *Towards Quantifying Depth and Size Perception in Virtual Environments*. Presence, 1995. **4**(1): p. 24-49.
8. Rolland, J.P., R.L. Holloway, and H. Fuchs. *Comparison of optical and video see-through head-mounted displays*. in *Proc. SPIE Vol. 2351-35*. 1994.
9. Edwards, E.K., J.P. Rolland, and K.P. Keller. *Video see-through design for merging of real and virtual environments*. in Proc. IEEE Virtual Reality International Symp. (VRAIS'93). 1993. Seattle.
10. Drascic, D. *Skill acquisition and task performance in teleoperation using monoscopic and stereoscopic video remote viewing*. in *Human Factors Society 35th Annual Meeting*. 1991. San Francisco, CA.
11. Meehan, J.W. and T.J. Triggs, *Magnification effects with imaging displays depend on scene content and viewing condition*. Human Factors, 1988. **30**(4): p. 487-494.
12. Woods, A., T. Docherty, and R. Koch. *Image distortions in stereoscopic video systems*. in *SPIE Volume 1915: Stereoscopic Displays and Applications IV*. 1993. San Jose.
13. Diner, D.B. *A new definition of orthostereopsis for 3-D television*. in *IEEE SMC International Conference on Systems, Man, and Cybernetics*. 1991.
14. Drascic, D. and J.J. Grodski. *Using Stereoscopic Video for Bomb Disposal Teleoperation*. in *SPIE 1915:*

Stereoscopic Displays and Applications IV. 1993. San Jose.

15. Drascic, D., *et al. ARGOS: stereoscopic video with stereoscopic graphics*. in *Video Proc. of INTERCHI'93: ACM Conference on Human Factors in Computing Systems*. 1993. Amsterdam, The Netherlands.
16. Diner, D. and D.H. Fender, *Human Engineering in Stereoscopic Viewing Devices*. Advances in Computer Vision and Machine Intelligence, ed. M.D. Levine. 1993, New York: Plenum Press.
17. Welch, R.B. and M.M. Cohen, *Adapting to variable prismatic displacement*, in *Pictorial Communication in Virtual and Real Environments*, S.R. Ellis, Editor. 1993, Taylor & Francis: London. p. 295-304.
18. Nemire, K. and S.R. Ellis. *Calibration and evaluation of virtual environment displays*. in *IEEE Computer Society & ACM/SIGGRAPH Research Frontiers in Virtual Reality*. 1993. San Jose.
19. Drascic, D. and P. Milgram. *Positioning Accuracy of a Virtual Stereographic Pointer in a Real Stereoscopic Video World*. in *SPIE Volume 1457: Stereoscopic Displays and Applications II*. 1991.
20. Utsumi, A., *et al. Investigation of Errors in Perception of Stereoscopically Presented Virtual Object Locations in Real Display Space*. in *Proc. of the Human Factors and Ergonomics Society 38th Annual Meeting*. 1994. Nashville.
21. Oman, C.M., *Sensory conflict in motion sickness: an Observer Theory approach*, in *Pictorial Communication in Virtual and Real Environments*, S.R. Ellis, Editor. 1993, Taylor & Francis: London. p. 362-376.
22. Rastogi, A., *Design of an interface for teleoperation in unstructured environments using augmented reality displays*, MASc Thesis 1996, University of Toronto, http://vered.rose.utoronto.ca/people/anu_dir/thesis/.
23. Boff, K.R. and J.E. Lincoln, *Engineering Data Compendium: Human Perception and Performance*. 1988, Wright-Patterson AFB, Ohio: AAMRL.
24. Baker, K. and H. Murray, *Striving for the Ultimate Image*, in *Optics and Photonics News*. 1993.
25. Panabaker, P.D., G.W. Harris, and W.C. Shaw. *Large Format Motion Pictures*. in *International Symposium on Three Dimensional Image Technology and Arts*. 1992. University of Tokyo, Seiken Symposium.
26. Merritt, J.O., *Personal Communication*, . 1996.
27. Diner, D.B. *Sub-pixel resolution in 3-D television for teleoperation*. in *Proceedings off the International Conference on Systems, Man, and Cybernetics*. 1991.
28. Drascic, D., *et al.*, *ARGOS: A Display System for Augmenting Reality*. ACM SIGGRAPH Technical Video Review: InterCHI '93 Conf on Human Factors in Computing Systems, (abstract appears in *Proceedings of InterCHI'93*, Amsterdam, p521), 1993. **88**(7).
29. Braunstein, M.L., *et al.*, *Recovering viewer-centered depth from disparity, occlusion, and velocity gradients*. Perception and Psychophysics, 1986. **40**(4): p. 216-224.
30. Butts, D.R.W. and D.F. McAllister, *Implementation of True 3D Cursors in Computer Graphics.* SPIE Volume 902: Three Dimensional Imaging and Remote Sensing Imaging, 1988: p. 74-84.
31. Noble, L. *A Single-Chip 3D Endoscope*. in *SPIE Volume 2653: Stereoscopic Displays and Applications VII*. 1996. San Jose.
32. Roscoe, S.N., *Judgments of size and distance with imaging displays*. Human Factors, 1984. **26**(6): p. 617-629.
33. Thouless, R.H., *Apparent size and distance in vision through a magnifying system*. British Journal of Psychology, 1968. **59**: p. 111-118.
34. Meehan, J.W. and T.J. Triggs, *Apparent size and distance in an imaging display*. Human Factors, 1992. **34**(3): p. 303-311.
35. Cegalis, J.A., *Prism distortion and accommodative change*. Perception and Psychophysics, 1973. **13**(3): p. 494-498.
36. Wade, N.J. and M. Swanston, *Visual Perception: An Introduction*. 1991, London: Routledge. 212.
37. Kersten, D. and G.E. Legge, *Convergence accommodation*. Journal of the Optical Society of America, 1983. **73**(3): p. 332-338.